CEUB

Introdução à Ciência de Dados

Medidas de Posição

MÉDIA ARITMÉTICA

Soma das observações, divididaa pelo número delas

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MEDIANA

Realização que ocupa a posição central da série de observações ordenadas.

- Se ímpar, valor central
- Se par, média dos dois valores centrais

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

MODA

Realização mais frequente do conjunto de valores observados

AMPLITUDE

Diferença entre o maior e o menor valor

Medidas de Posição

Calcule

$$7, 11, 11, 15, 20, 20, 28$$

Média

Moda

Mediana

Amplitude

Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representatividade posição central, não informa sobre a variabilidade

Cinco grupos de alunos submeteram-se a um teste, obtendo as seguintes notas:

Grupo A: **3, 4, 5, 6, 7**

Grupo B: **1, 3, 5, 7, 9**

Grupo C: **5, 5, 5, 5, 5**

Grupo D: **3, 5, 5, 7**

Grupo E: **3, 5, 5, 6, 6**

Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representatividade posição central, não informa sobre a variabilidade

Cinco grupos de alunos submeteram-se a um teste, obtendo as seguintes notas:

Grupo A: **3, 4, 5, 6, 7**

Grupo B: **1, 3, 5, 7, 9**

Grupo C: **5, 5, 5, 5, 5**

Grupo D: **3, 5, 5, 7**

Grupo E: **3, 5, 5, 6, 6**

O que conseguimos informar sobre os grupos?

Medidas de Dispersão

Dispersão em torno da média

Desvio médio

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Variância

População

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Amostra

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Medidas de Dispersão

Exemplor anterior

VALORES	MÉDIA	DESVIO	DESVIO QUAD.
7			
11			
11			
15			
20			
20			
28			

Medidas de Dispersão

Cinco grupos de alunos submeteram-se a um teste, obtendo as seguintes notas:

Grupo A: **3, 4, 5, 6, 7**

Grupo B: **1, 3, 5, 7, 9**

Grupo C: **5, 5, 5, 5, 5**

Grupo D: **3, 5, 5, 7**

Grupo E: **3, 5, 5, 6, 6**

O que conseguimos informar sobre os grupos?

Quantis Empíricos

Quantil de ordem p ou p-quantil, indicado por q(p), é uma medida tal que, 100p% das observações sejam menores do que q(p).

25%	25%	25%	25%

Quantis Empíricos

Quantil de ordem p ou p-quantil, indicado por q(p), é uma medida tal que, 100p% das observações sejam menores do que q(p).

Q1 **Q2** **Q3**

25%	25%	25%	25%

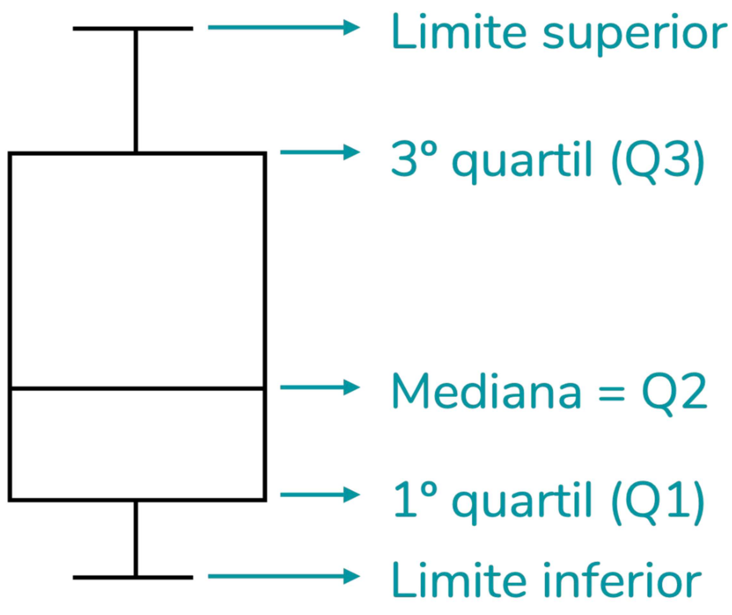
Quantis Empíricos

Quantil de ordem p ou p-quantil, indicado por q(p), é uma medida tal que, 100p% das observações sejam menores do que q(p).

	Q1		Q2		Q3	
25%		25%		25%		25%

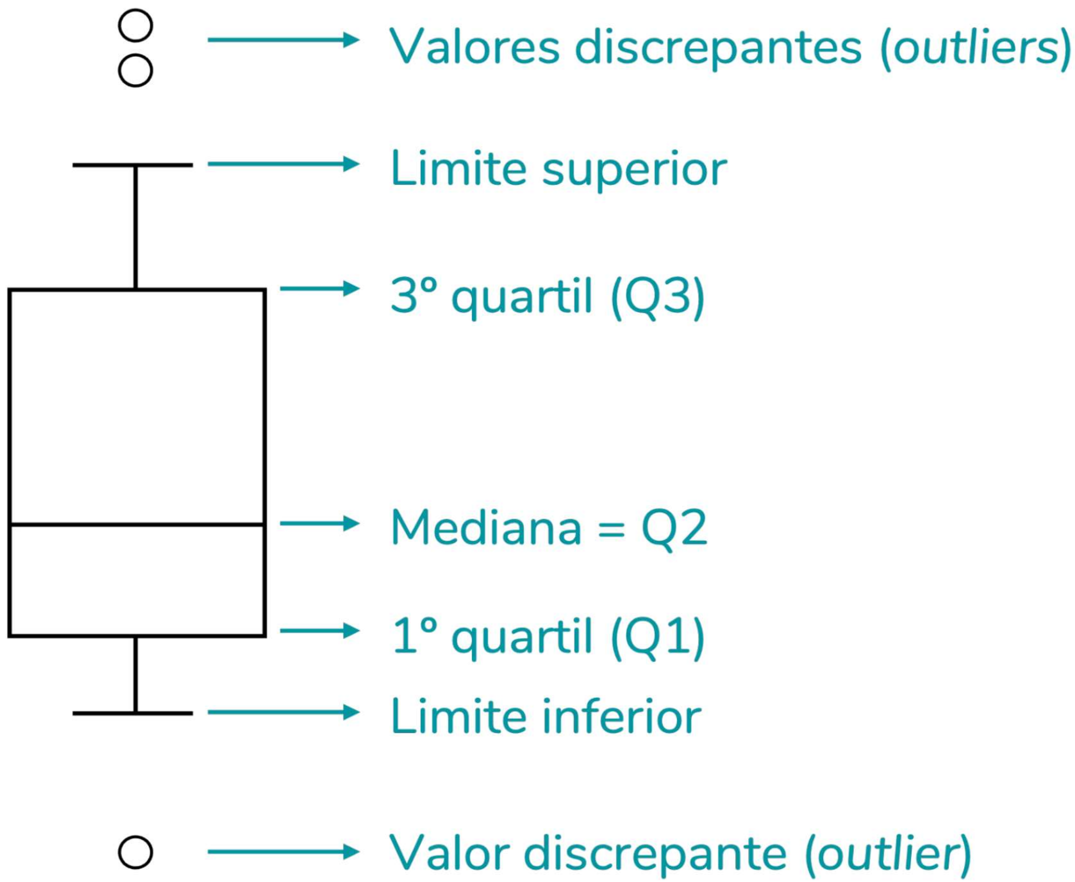
Intervalo Interquartil

Box-plot

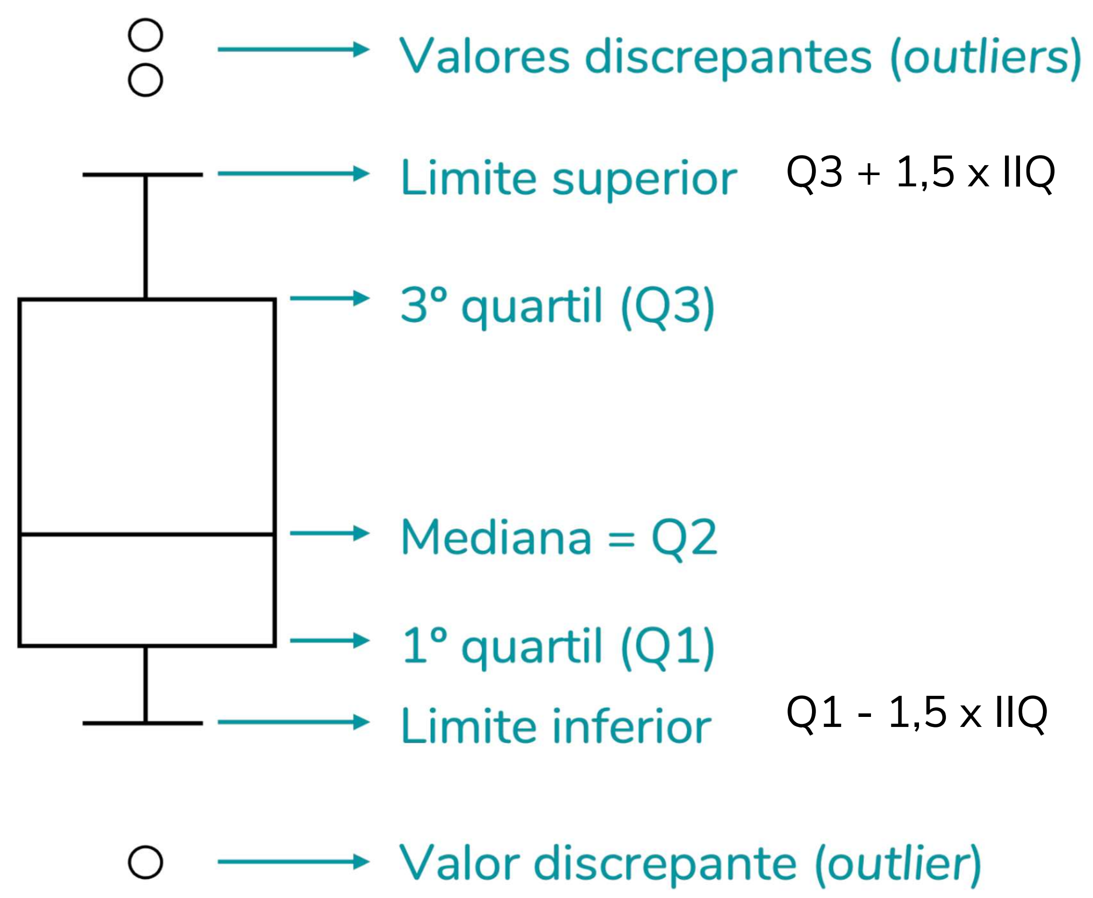


Limite superior
3º quartil (Q3)
Mediana = Q2
1º quartil (Q1)
Limite inferior

Box-plot


Valores discrepantes (*outliers*)
Limite superior
3º quartil (Q3)
Mediana = Q2
1º quartil (Q1)
Limite inferior
Valor discrepante (*outlier*)

Box-plot



Valores discrepantes (*outliers*)
Limite superior
Q3 + 1,5 x IIQ
3° quartil (Q3)
Mediana = Q2
1° quartil (Q1)
Limite inferior
Q1 - 1,5 x IIQ
Valor discrepante (*outlier*)

Box-plot

Calcule

22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79

Limite Inferior

Primeiro Quartil

Mediana

Terceiro Quartil

Limite Superior

R

Lim. Sup: 47.25

Q3: 33

Mediana: 25.5

Q1: 23.5

Lim. Inf: 22

IQR: 9.5

22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79

R

Lim. Sup: 47.25

Q3: 33

Mediana: 25.5

Q1: 23.5

Lim. Inf: 22

IQR: 9.5

22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79

```r
v1 <- c(22, 22, 23,
        24, 25, 25,
        26, 26, 27,
        39, 59, 79)

box_v1 <- boxplot(v1)
```

R

Lim. Sup: 47.25

Q3: 33

Mediana: 25.5

Q1: 23.5

Lim. Inf: 22

IQR: 9.5

22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79

```
> box_v1$stats
     [,1]
[1,] 22.0
[2,] 23.5
[3,] 25.5
[4,] 33.0
[5,] 39.0
```

Stack Overflow

3 Answers

Sorted by: Highest score (default)

11

The values of the box are called hinges and may coincide with the quartiles (as calculated by `quantile(x, c(0.25, .075))`), but are calculated differently.

From `?boxplot.stats`:

> The two ‘hinges’ are versions of the first and third quartile, i.e., close to quantile(x, c(1,3)/4). The hinges equal the quartiles for odd n (where n <- length(x)) and differ for even n. Whereas the quartiles only equal observations for n %% 4 == 1 (n = 1 mod 4), the hinges do so additionally for n %% 4 == 2 (n = 2 mod 4), and are in the middle of two observations otherwise.

To see that the values coincide with an odd number of observations, try the following code:

```
set.seed(1234)
x <- rnorm(9)

boxplot(x)
abline(h=quantile(x, c(0.25, 0.75)), col="red")
```

?quantile

Types

`quantile` returns estimates of underlying distribution quantiles based on one or two order statistics from the supplied elements in `x` at probabilities in `probs`. One of the nine quantile algorithms discussed in Hyndman and Fan (1996), selected by `type`, is employed.

All sample quantiles are defined as weighted averages of consecutive order statistics. Sample quantiles of type i are defined by:

$$Q_i(p) = (1-\gamma)x_j + \gamma x_{j+1}$$

where $1 \leq i \leq 9$, $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$, x_j is the jth order statistic, n is the sample size, the value of γ is a function of $j = \lfloor np + m \rfloor$ and $g = np + m - j$, and m is a constant determined by the sample quantile type.

STATISTICAL COMPUTING

This department includes the two sections *New Development in Statistical Computing* and *Statistical Computing Software Reviews*; suitable contents for each of these sections are described under the respective section heading. Articles submitted for the department, outside the two sections, should not be highly technical and should be relevant to the teaching or practice of statistical computing.

Sample Quantiles in Statistical Packages

Rob J. Hyndman and Yanan Fan

There are a large number of different definitions used for sample quantiles in statistical computer packages. Often within the same package one definition will be used to compute a quantile explicitly, while other definitions may be used when producing a boxplot, a probability plot, or a QQ plot. We compare the most commonly implemented sample quantile definitions by writing them in a common notation and investigating their motivation and some of their properties. We argue that there is a need to adopt a standard definition for sample quantiles so that the same answers are produced by different packages and within each package. We conclude by recommending that the median-unbiased estimator be used because it has most of the desirable properties of a quantile estimator and can be defined independently of the underlying distribution.

KEY WORDS: Percentiles; Quartiles; Sample quantiles; Statistical computer packages.

1. INTRODUCTION

The quantile of a distribution is defined as

$$Q(p) = F^{-1}(p) = \inf\{x\colon F(x) \geq p\}, \qquad 0 < p < 1,$$

where $F(x)$ is the distribution function. Sample quantiles provide nonparametric estimators of their population counterparts based on a set of independent observations

can be written as

$$\hat{Q}_i(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}$$

$$\text{where} \quad \frac{j-m}{n} \leq p < \frac{j-m+1}{n} \qquad (1)$$

for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$. The value of γ is a function of $j = \lfloor pn + m \rfloor$ and $g = pn + m - j$. Here, $\lfloor u \rfloor$ denotes the largest integer not greater than u; later we shall use $\lceil u \rceil$ to denote the smallest integer not less than u.

We consider estimators of the form (1), including some that are not found in statistical packages. There have been several other nonparametric quantile estimators proposed that are not of the form (1) (e.g., Harrell and Davis 1982; Sheather and Marron 1990), but these are not implemented in widely available packages and so are not considered here. We also exclude sample quantiles that are not defined for all p including hinges and other letter values (Hoaglin 1983) and related methods (Freund and Perles 1987).

A closely related problem is the selection of plotting position in a quantile plot in which $X_{(k)}$ is plotted against p_k or in a quantile–quantile plot in which $X_{(k)}$ is plotted against $G^{-1}(p_k)$ where G is a distribution function. Various rules for p_k have been suggested (see Cunnane 1978; Harter 1984; Kimball 1960; Mage 1982). Each plotting rule corresponds to a sample quantile definition by defining $\hat{Q}_i(p_k) = X_{(k)}$ and using linear interpolation for $p \neq p_k$. However, the criteria by which a plotting position is chosen (e.g., the five postulates of Gumbel 1958, pp. 32–34 or the three purposes of Kimball 1960) may be quite different from the criteria for choosing a good sample quantile definition.

We compare sample quantile definitions of the form (1)

?boxplot.stats

Details

The two 'hinges' are versions of the first and third quartile, i.e., close to `quantile(x, c(1,3)/4)`. The hinges equal the quartiles for odd n (where `n <- length(x)`) and differ for even n. Whereas the quartiles only equal observations for `n %% 4 == 1` ($n \equiv 1 \bmod 4$), the hinges do so *additionally* for `n %% 4 == 2` ($n \equiv 2 \bmod 4$), and are in the middle of two observations otherwise.

The notches (if requested) extend to `+/-1.58 IQR/sqrt(n)`. This seems to be based on the same calculations as the formula with 1.57 in Chambers *et al* (1983, p. 62), given in McGill *et al* (1978, p. 16). They are based on asymptotic normality of the median and roughly equal sample sizes for the two medians being compared, and are said to be rather insensitive to the underlying distributions of the samples. The idea appears to be to give roughly a 95% confidence interval for the difference in two medians.

?boxplot.stats

Details

The two 'hinges' are versions of the first and third quartile, i.e., close to `quantile(x, c(1,3)/4)`. The hinges equal the quartiles for odd n (where `n <- length(x)`) and differ for even n. Whereas the quartiles only equal observations for `n %% 4 == 1` ($n \equiv 1 \bmod 4$), the hinges do so *additionally* for `n %% 4 == 2` ($n \equiv 2 \bmod 4$), and are in the middle of two observations otherwise.

The notches (if requested) extend to `+/-1.58 IQR/sqrt(n)`. This seems to be based on the same calculations as the formula with 1.57 in Chambers *et al* (1983, p. 62), given in McGill *et al* (1978, p. 16). They are based on asymptotic normality of the median and roughly equal sample sizes for the two medians being compared, and are said to be rather insensitive to the underlying distributions of the samples. The idea appears to be to give roughly a 95% confidence interval for the difference in two medians.

R

22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79

Refaça o exercício excluindo o último valor

Muitos Dados

Entendo o processo com 10 observações,

Podemos expandir a ideia, trabalhando com muitas observações

RH

Análise Curricular

...

Entrevista técnica

Entrevistas em Ciência de Dados

Estudo de caso envolvendo área de atuação da empresa/cargo

- **Dados mascarados ou fictícios**
- **Fidedignos aos dados reais**

https://github.com/ifood/ifood-data-analyst-case

main
1 branch
0 tags
Go to file
Code
1uisinaugusto Add files via upload
ab2e0ab on Mar 2, 2022
2 commits
.gitignore
Initial commit
last year
LICENSE
Initial commit
last year
README.md
Initial commit
last year
Retail Company Case.pdf
Add files via upload
last year
retail_case_data.csv
Add files via upload
last year
README.md
ifood-data-analyst-case
repositório destinado ao case de contratação do time de data & analytics

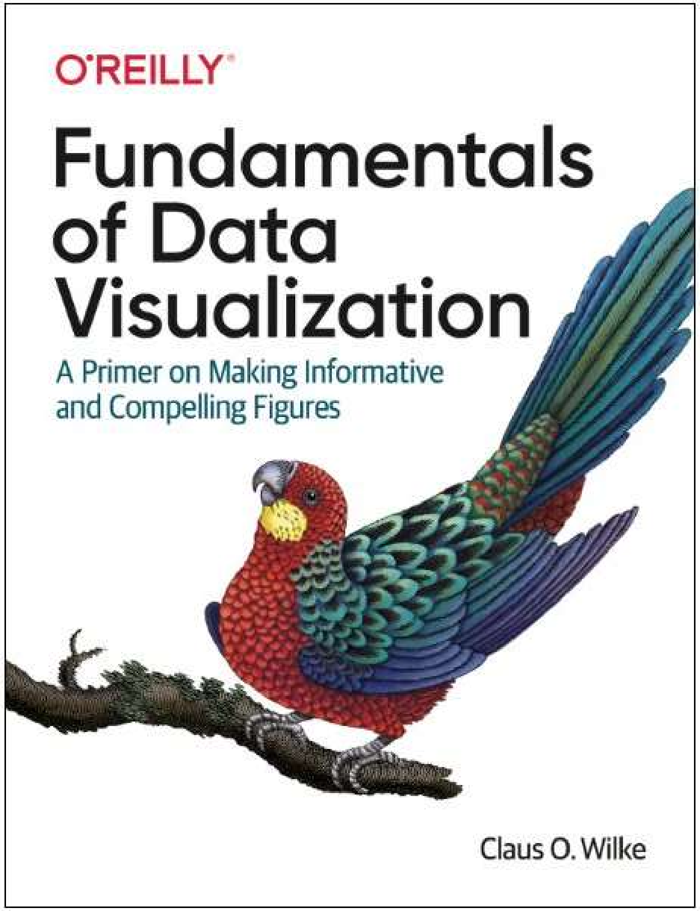
Visualização de dados

Porque “olhar” para os dados?

Gráficos

Transmitir a mensagem presente nos dados/valores computados

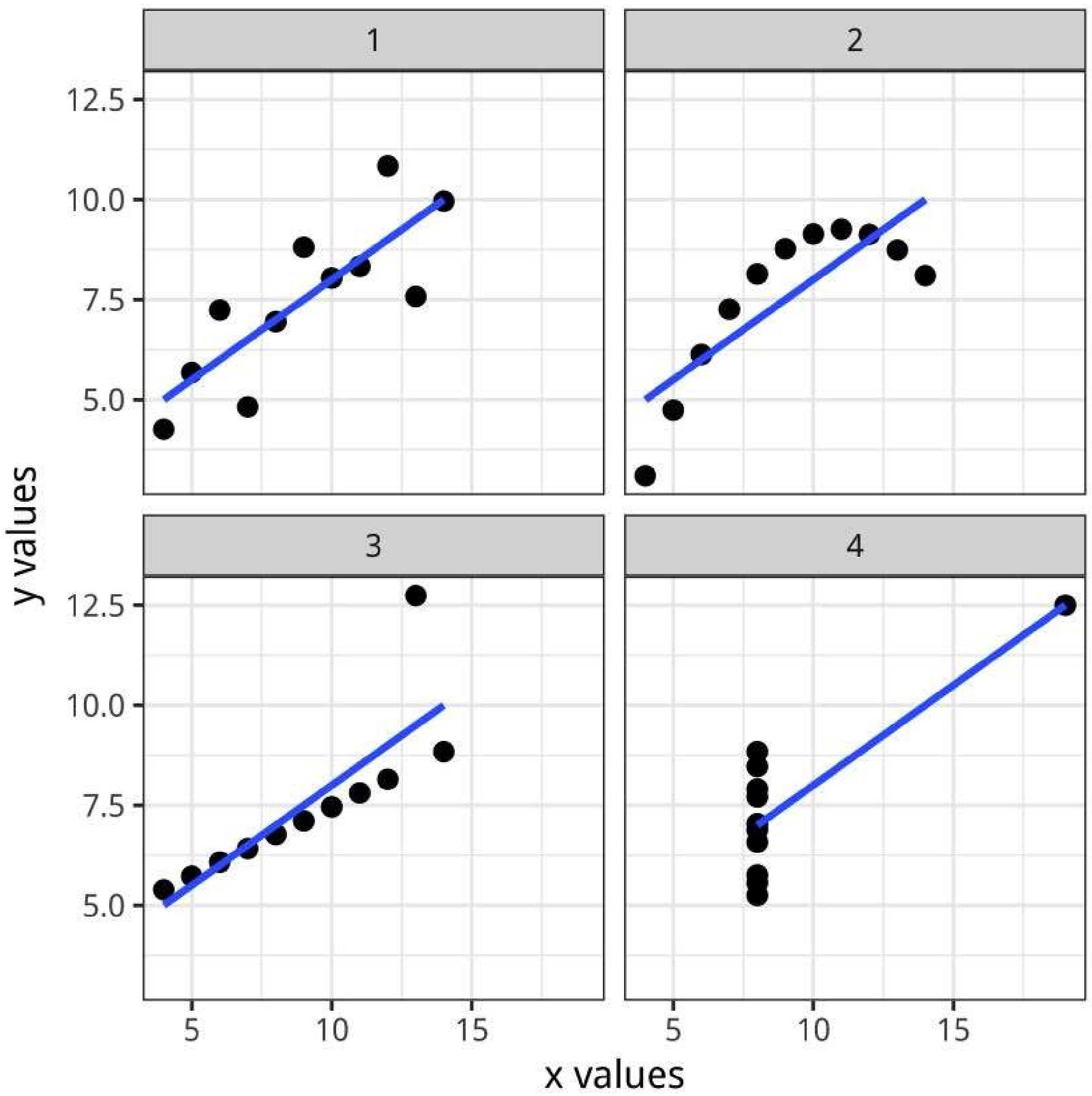
Explorar e investigar a estrutura dos seus dados



“Figures will typically carry the weight of your arguments”[1]

Dados de Anscombe

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03


1
2
3
4
12.5
10.0
7.5
5.0
5
10
15
x values
y values

Visualização de dados

gnuplot, Xfig, Mathematica, Matlab,

Matplotlib, seaborn, plotly

base R, ggplot2,

. . .

Constante mudança de softwares

Visualização de dados

Arte certa sem a ciência errada

Mensagem clara e convincente

Visualmente agradável

Visualização de dados

Gráfico Feio

Figura com problemas estéticos, mas é clara e informativa

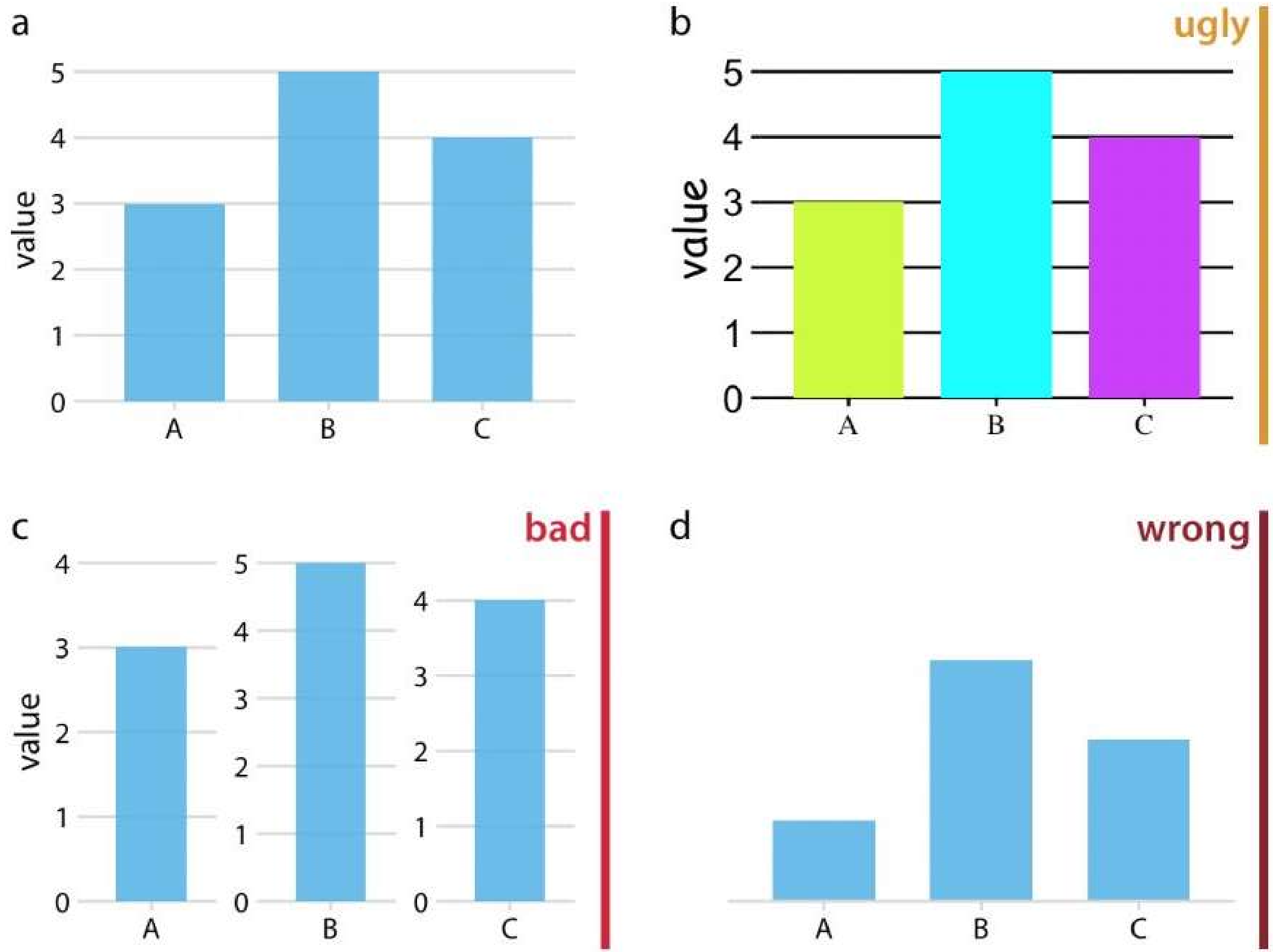
Gráfico Ruim

Figura com problemas relacionados à percepção; mensagem confusa, informação não clara

Gráfico Errado

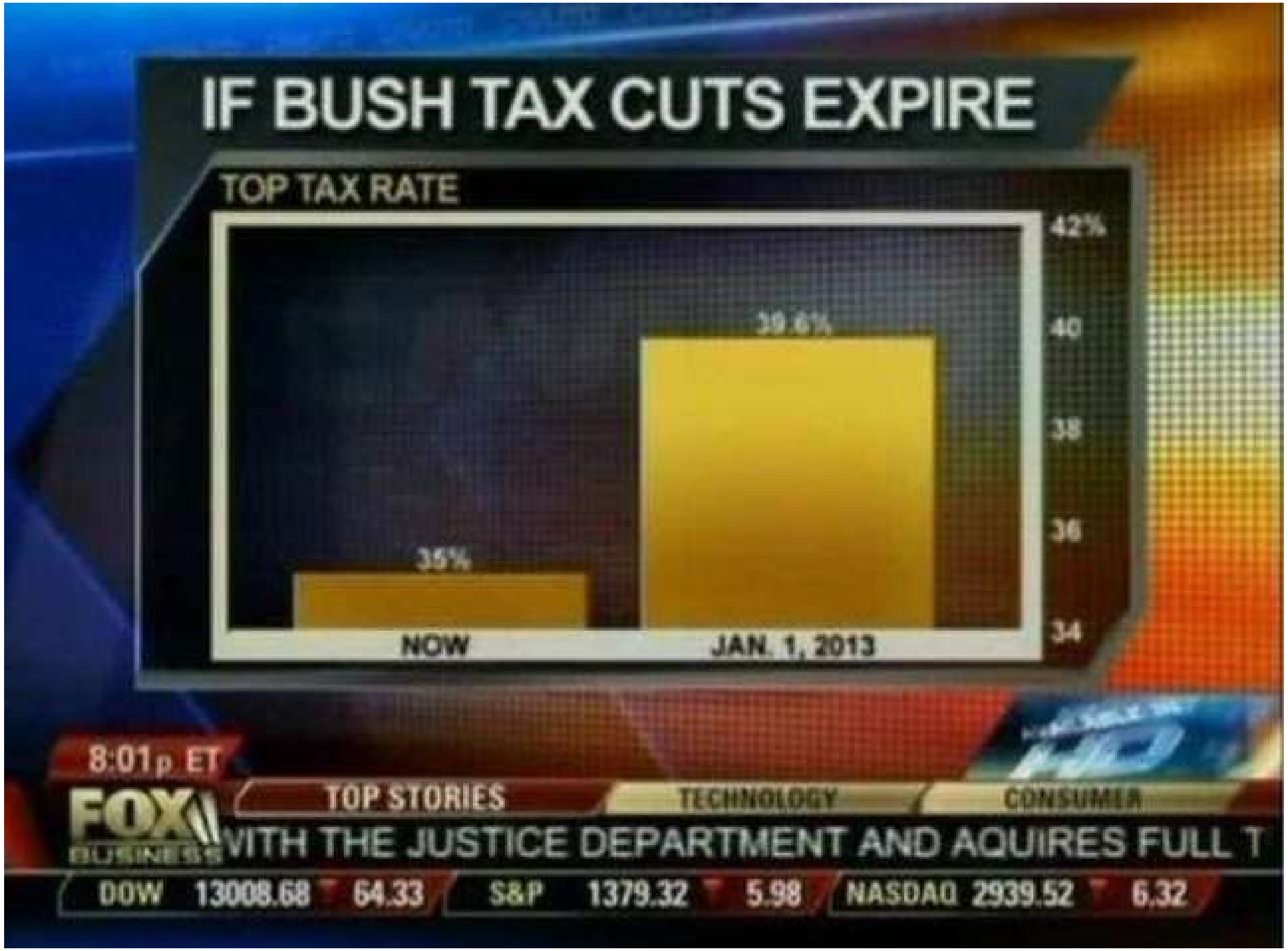
Figura matematicamente errada. Passa uma informação incorreta.

Visualização de dados

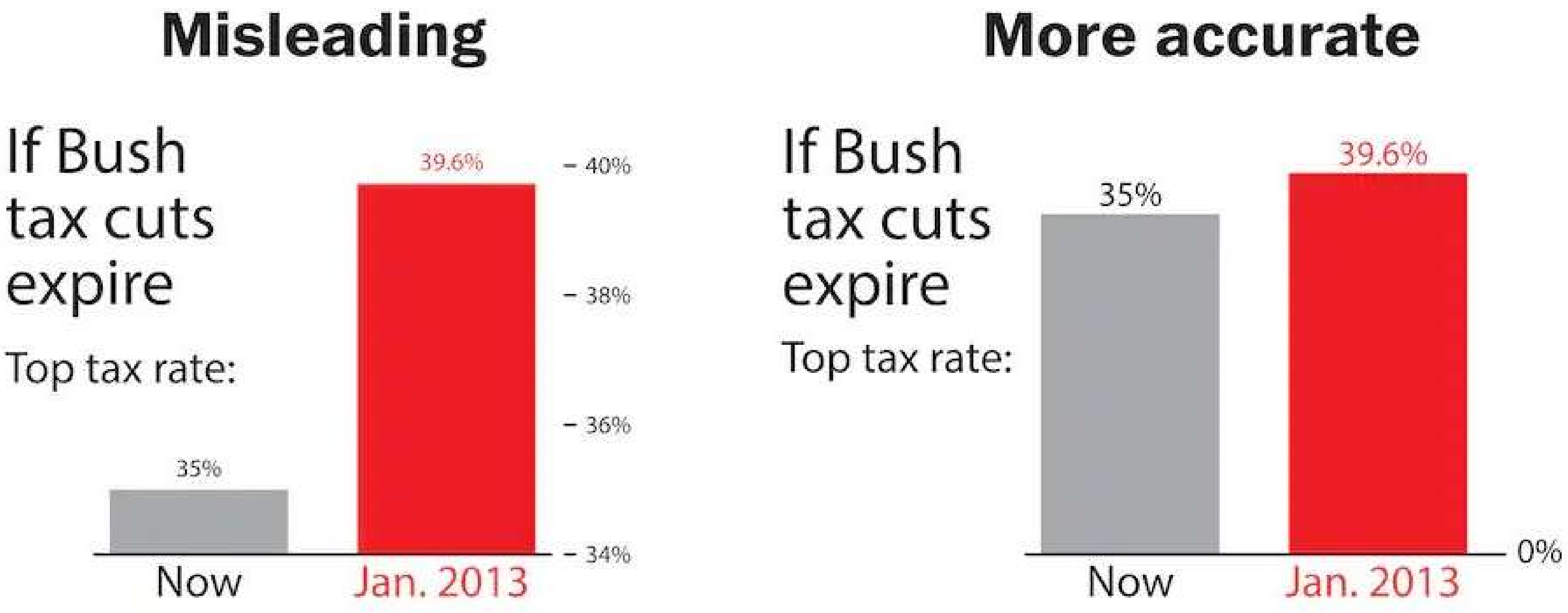



Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Africa
Americas
Asia
Europe
Oceania
Life Expectancy: 2007

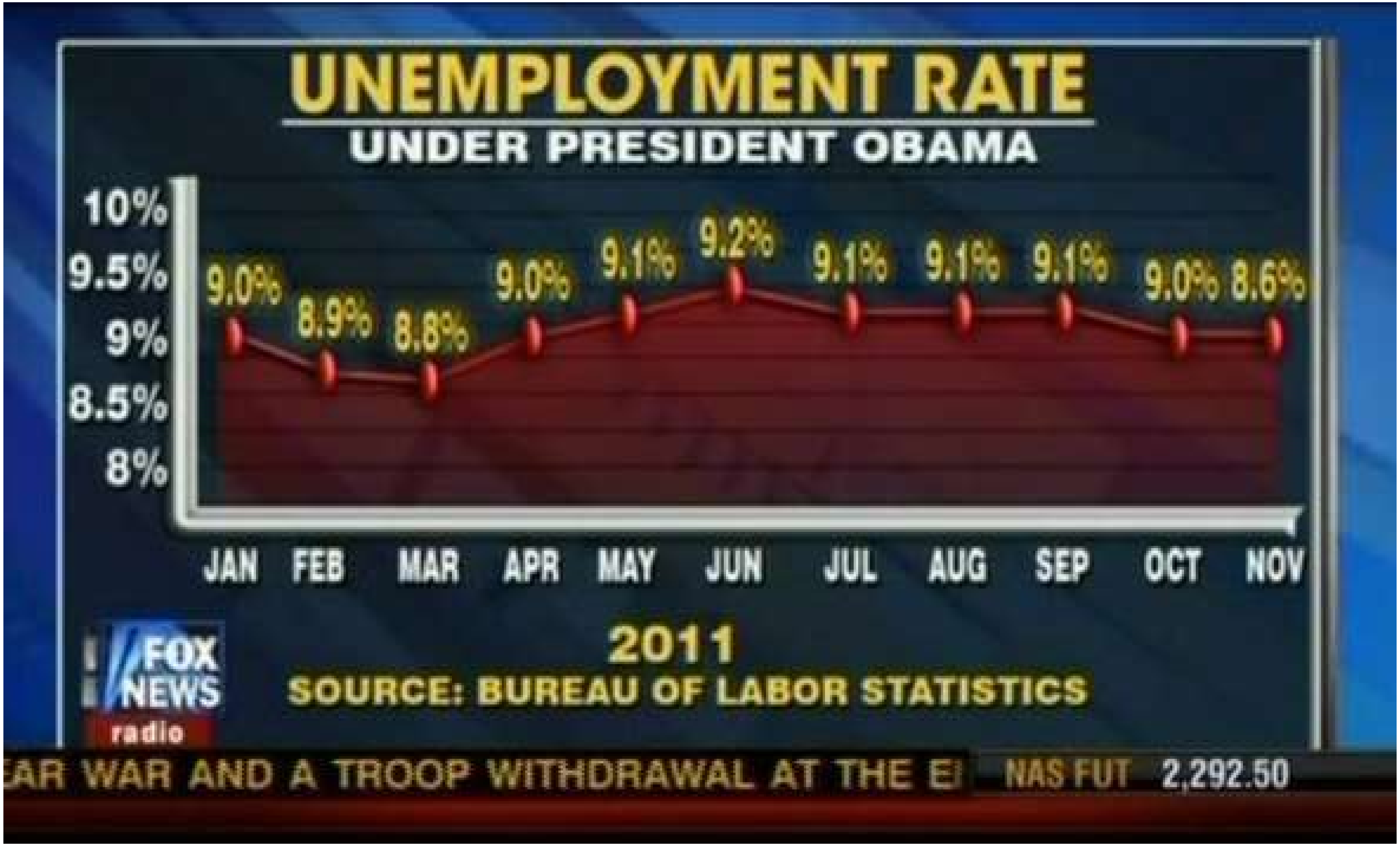
Menções Honrosas



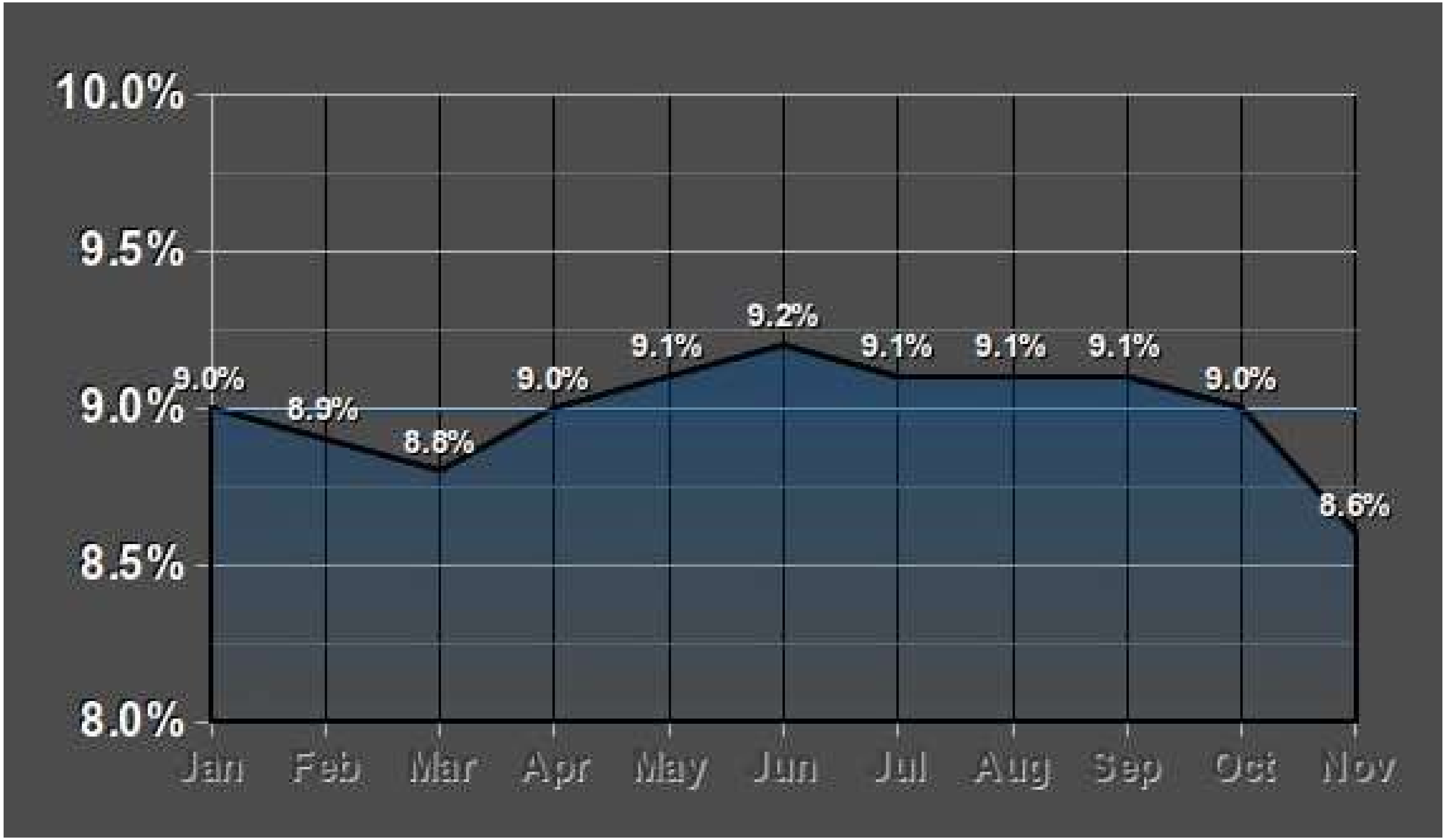
https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/



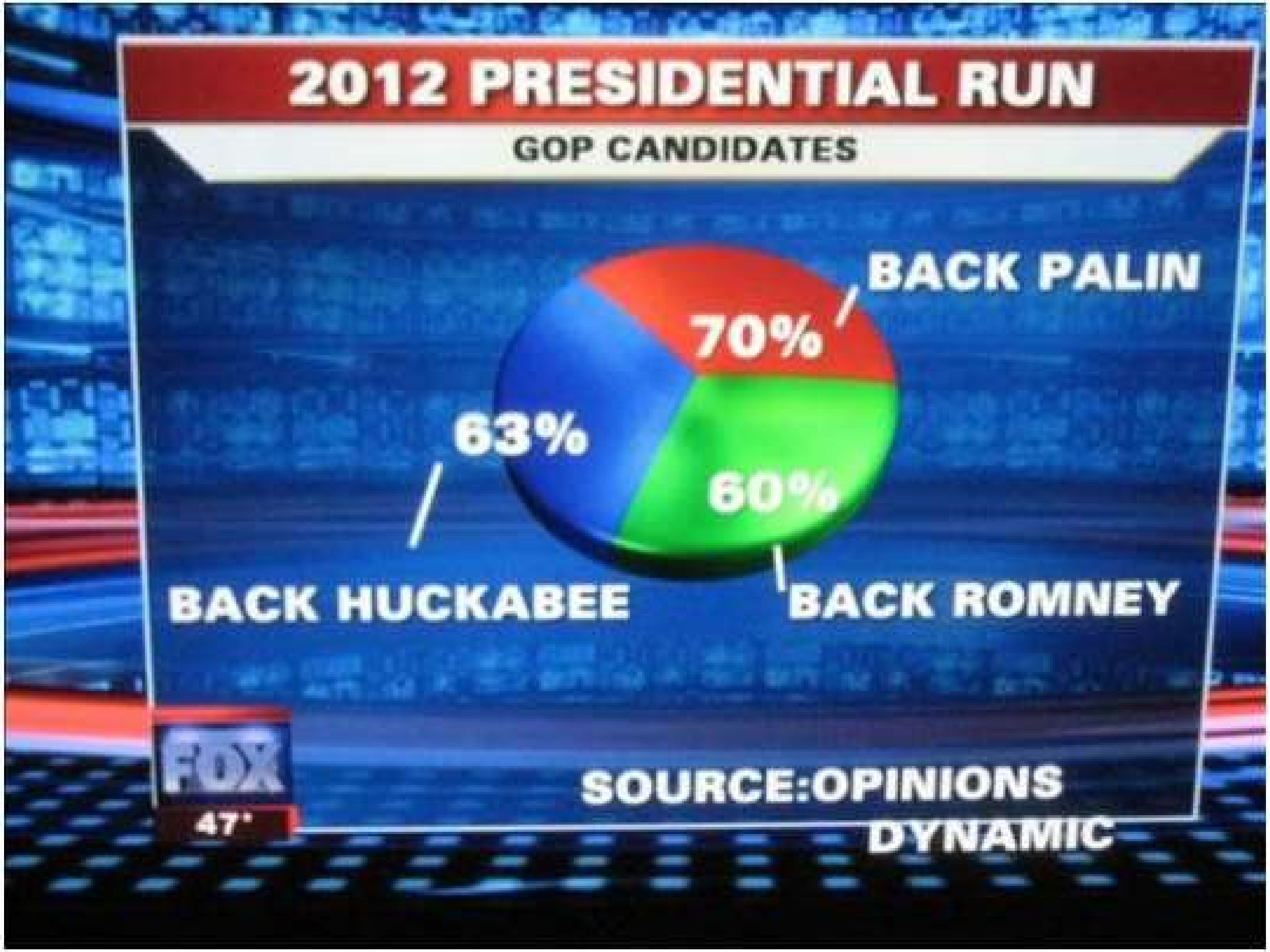
Fonte: https://www.washingtonpost.com/business/2019/10/14/youve-been-reading-charts-wrong-heres-how-pro-does-it/



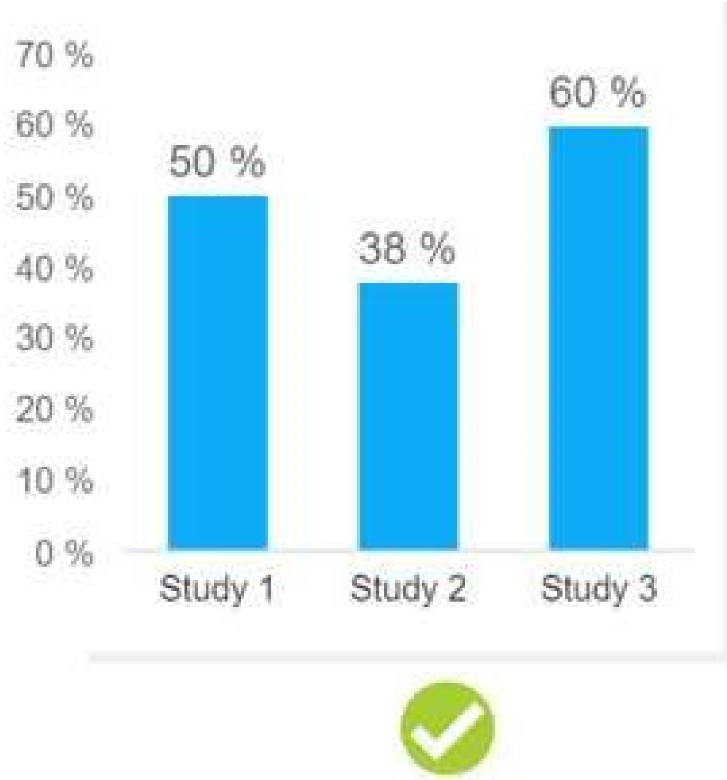
UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
2011
SOURCE: BUREAU OF LABOR STATISTICS
FOX NEWS radio
EAR WAR AND A TROOP WITHDRAWAL AT THE E
NAS FUT 2,292.50



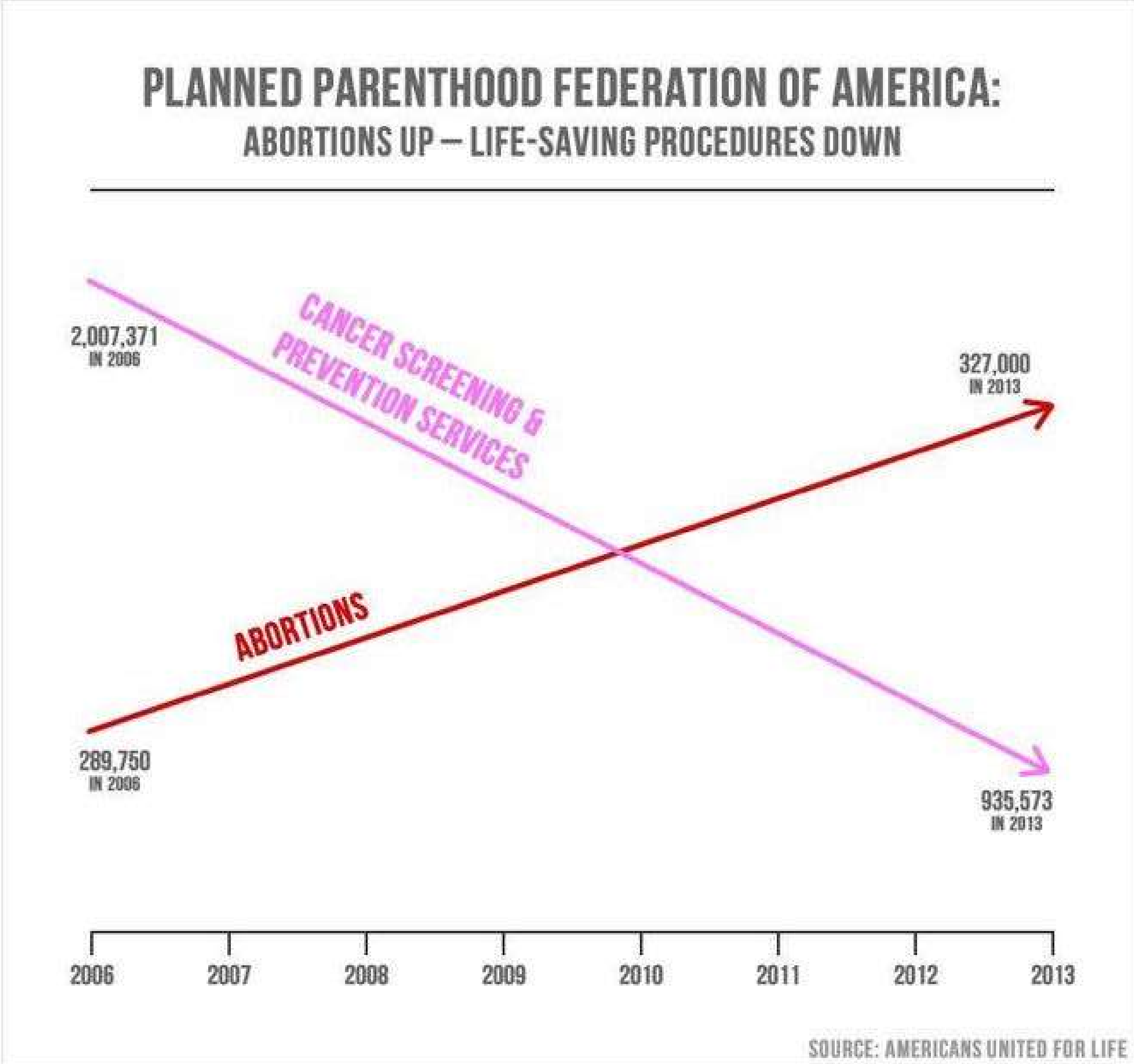
Fonte: http://freethoughtblogs.com/lousycanuck/2011/12/14/im-better-at-graphs-than-fox-news/



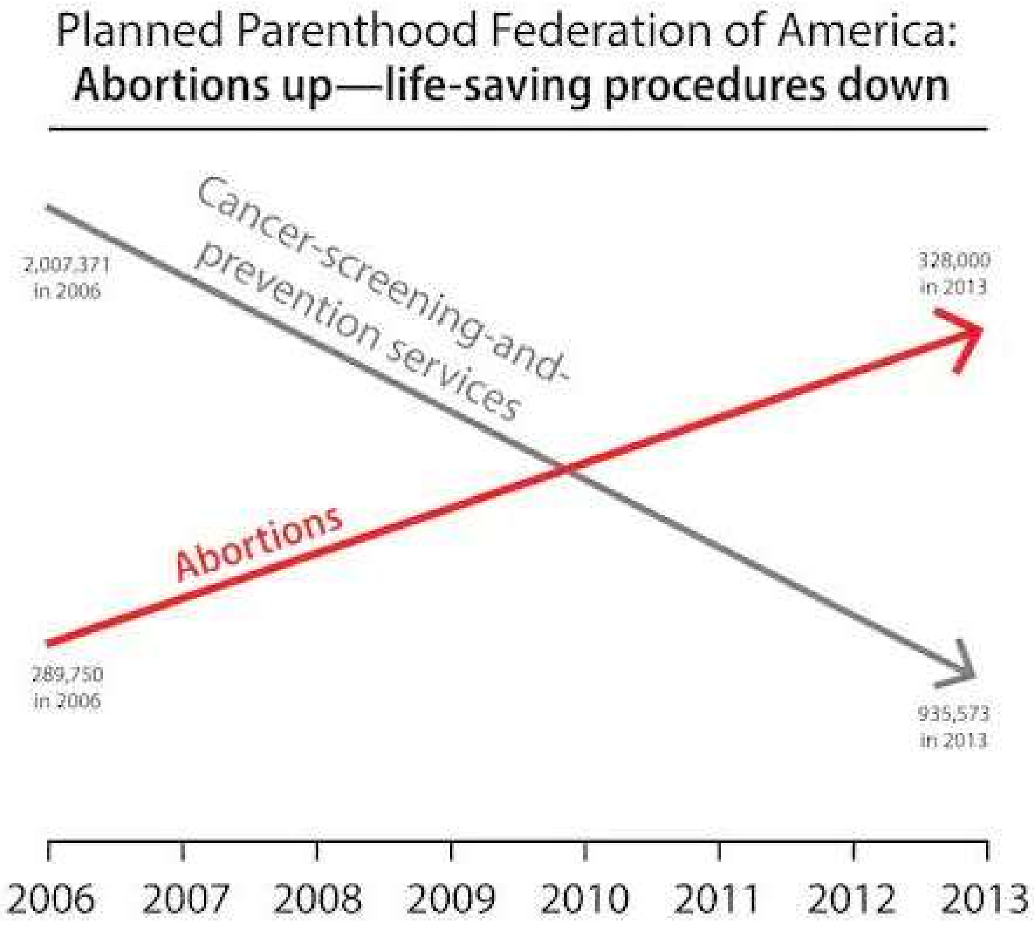
Fonte: http://freethoughtblogs.com/lousycanuck/2011/12/14/im-better-at-graphs-than-fox-news/



Fonte: https://www.datapine.com/blog/misleading-data-visualization-examples/

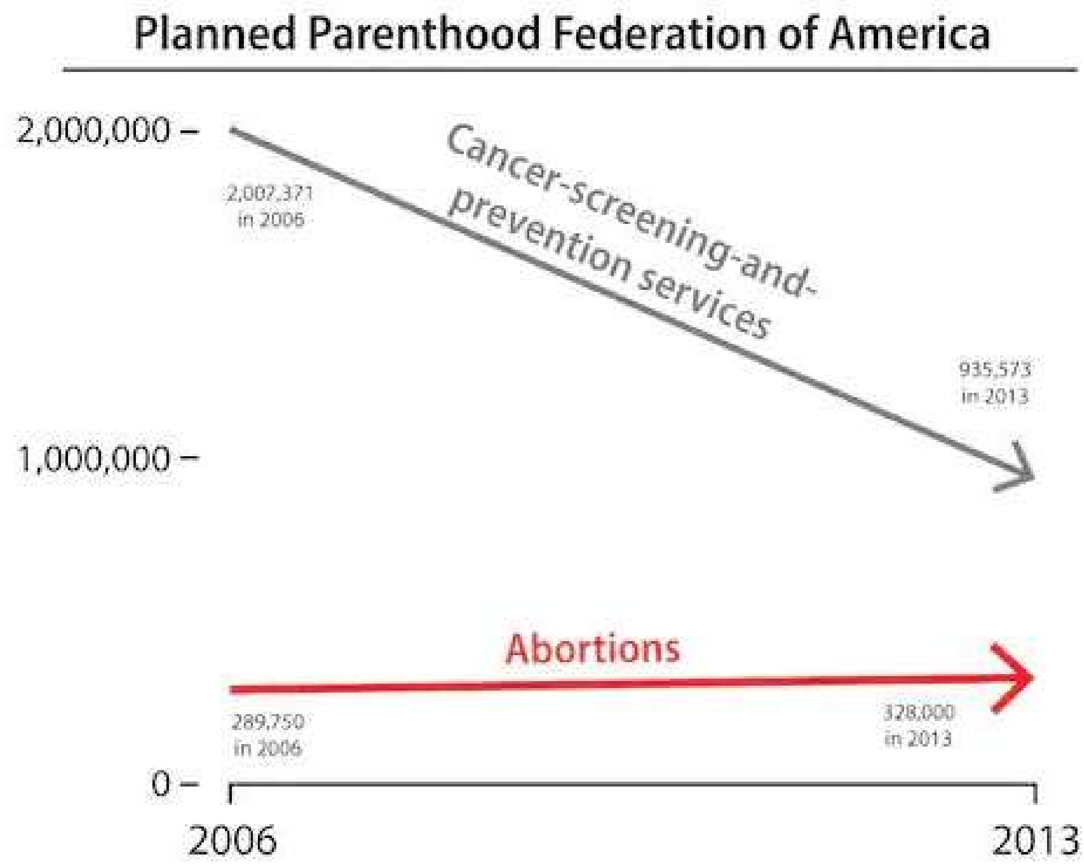

PLANNED PARENTHOOD FEDERATION OF AMERICA:
ABORTIONS UP — LIFE-SAVING PROCEDURES DOWN
2,007,371
IN 2006
CANCER SCREENING &
PREVENTION SERVICES
327,000
IN 2013
ABORTIONS
289,750
IN 2006
935,573
IN 2013
2006
2007
2008
2009
2010
2011
2012
2013
SOURCE: AMERICANS UNITED FOR LIFE

Misleading



(Source: Americans United for Life)

More accurate

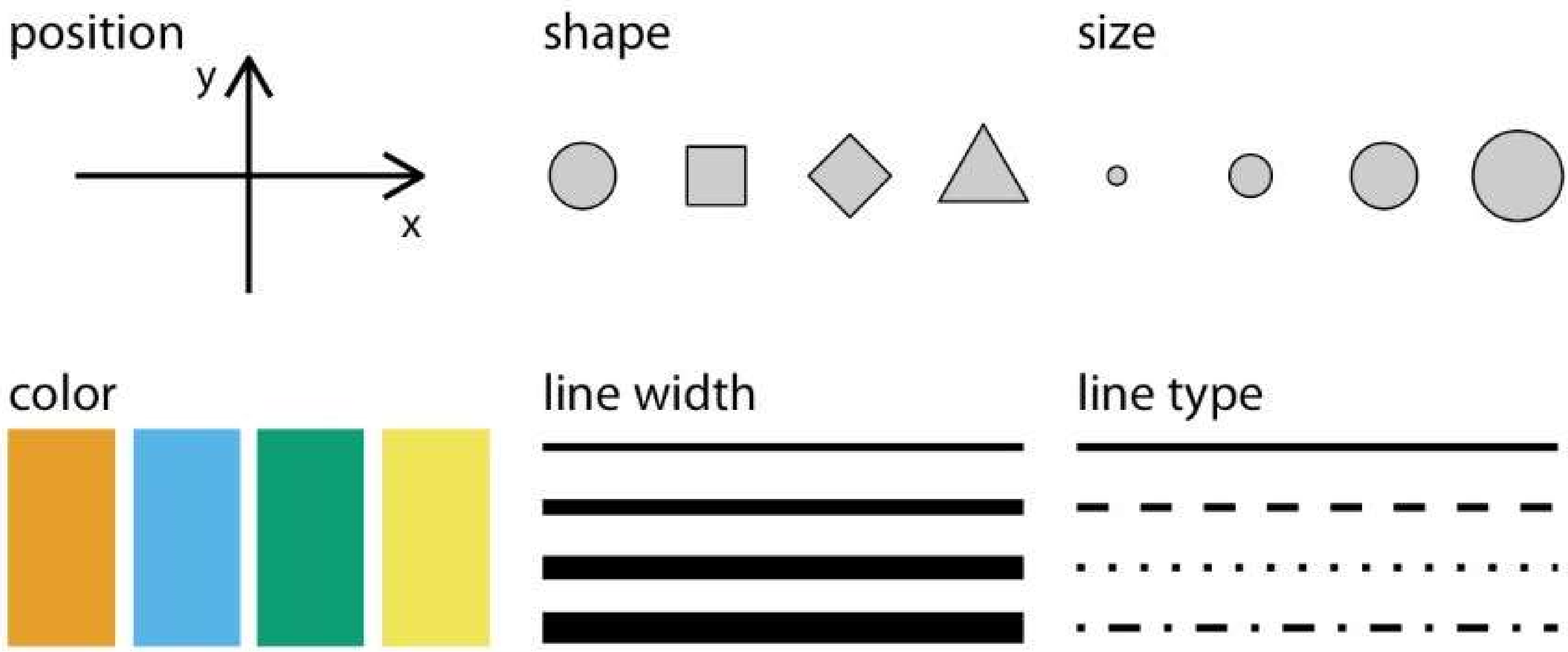


COMO CRIAR GRÀFICOS?

Mapeamento Estético

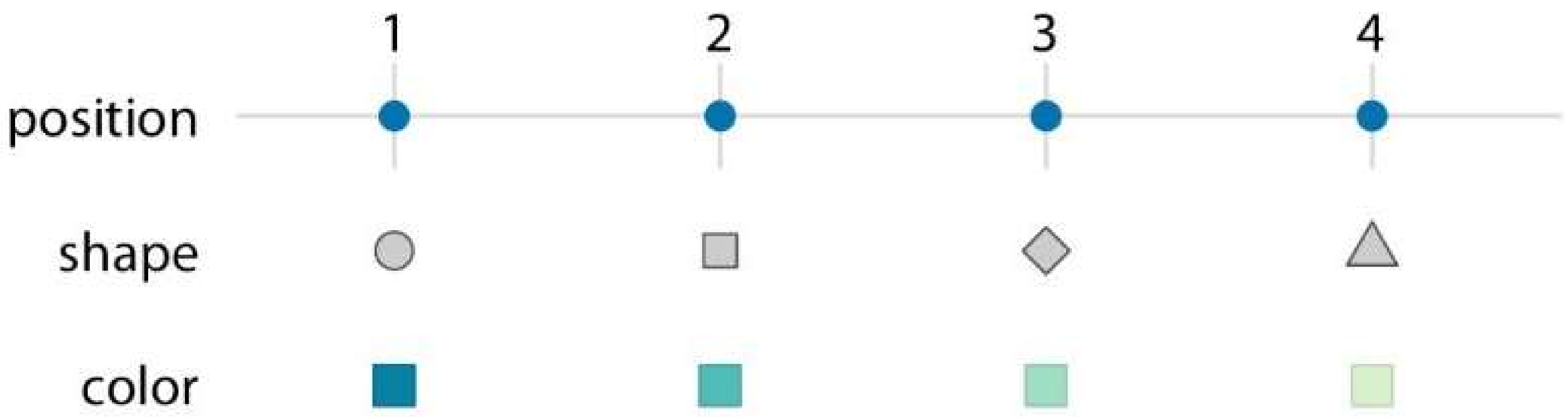
Todas as visualzações de dados mapeiam valores de dados em camadas quantificáveis do gráfico. Tais camadas, são as camadas estéticas.

aesthetics



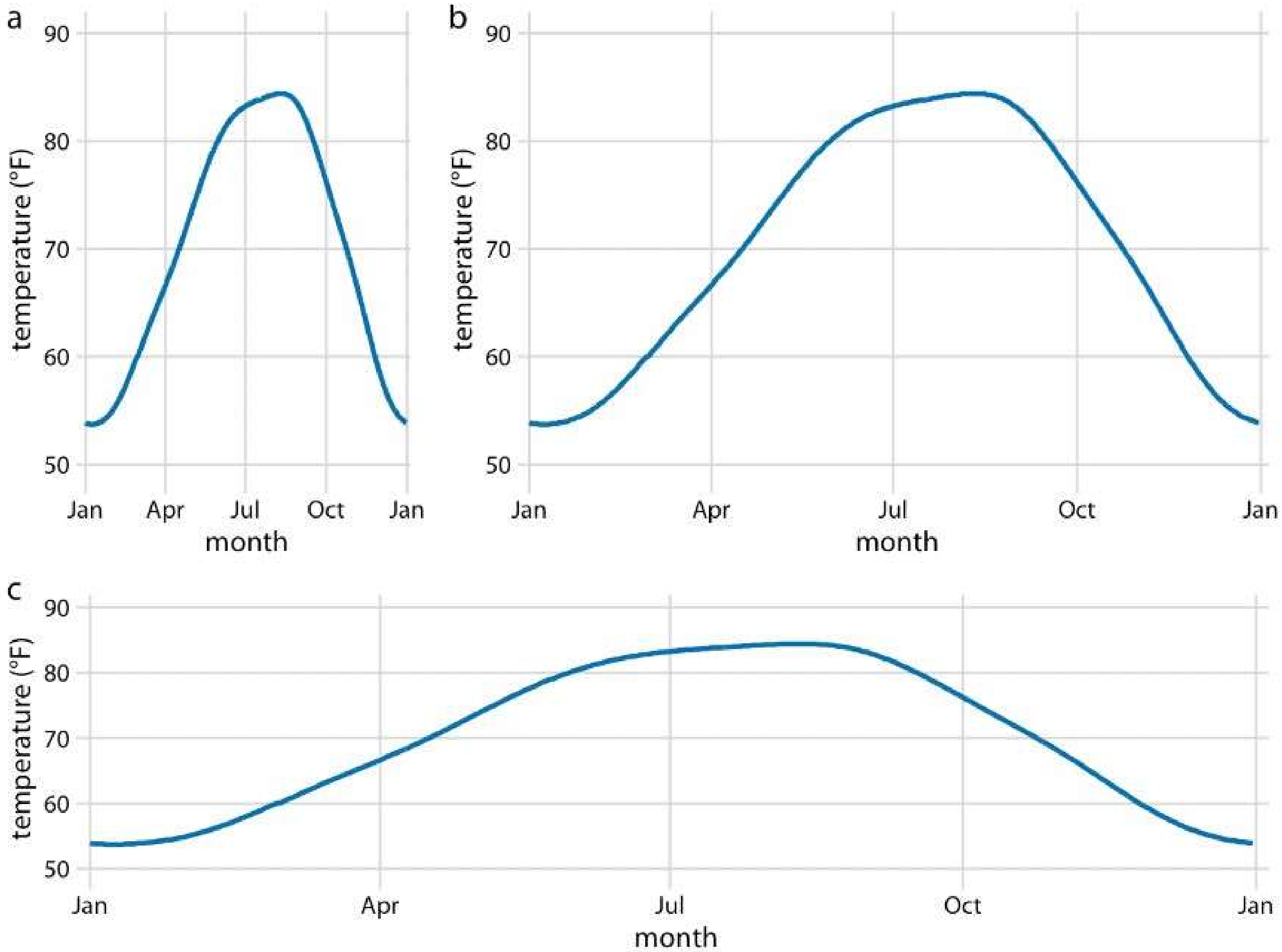
Fonte: Fundamentals of Data Visualization - Claus O. Wilke

scales

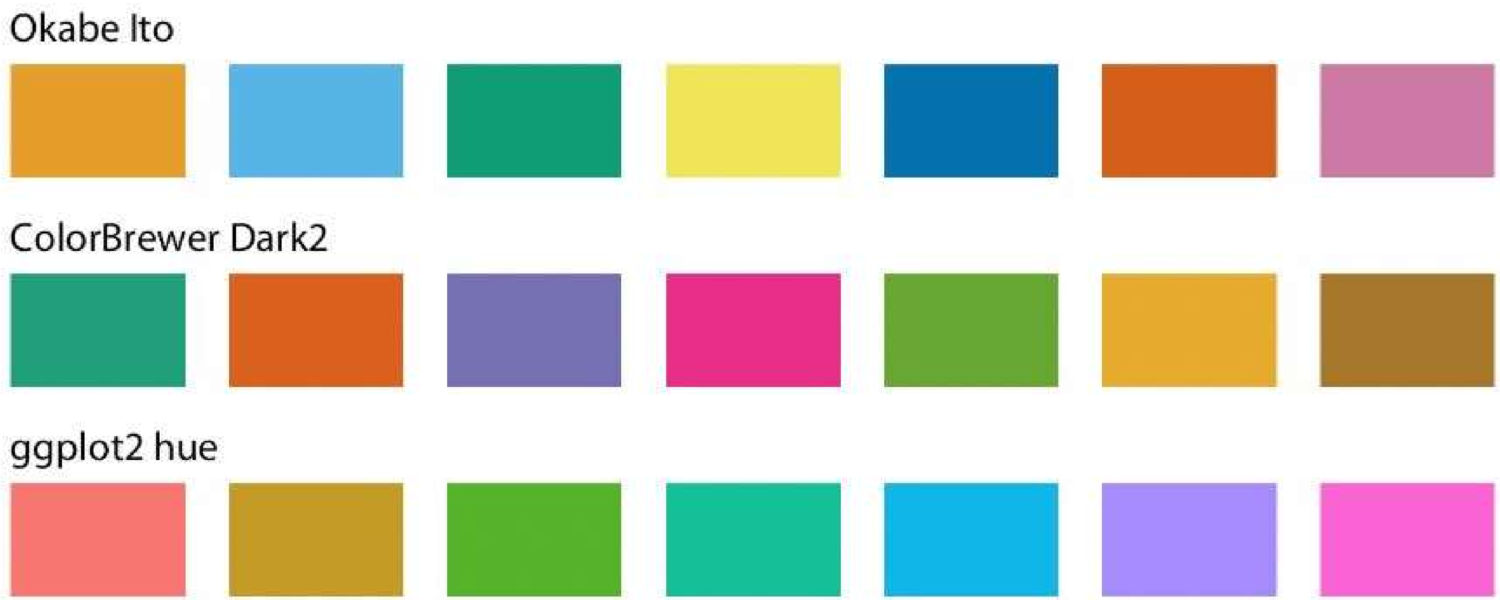


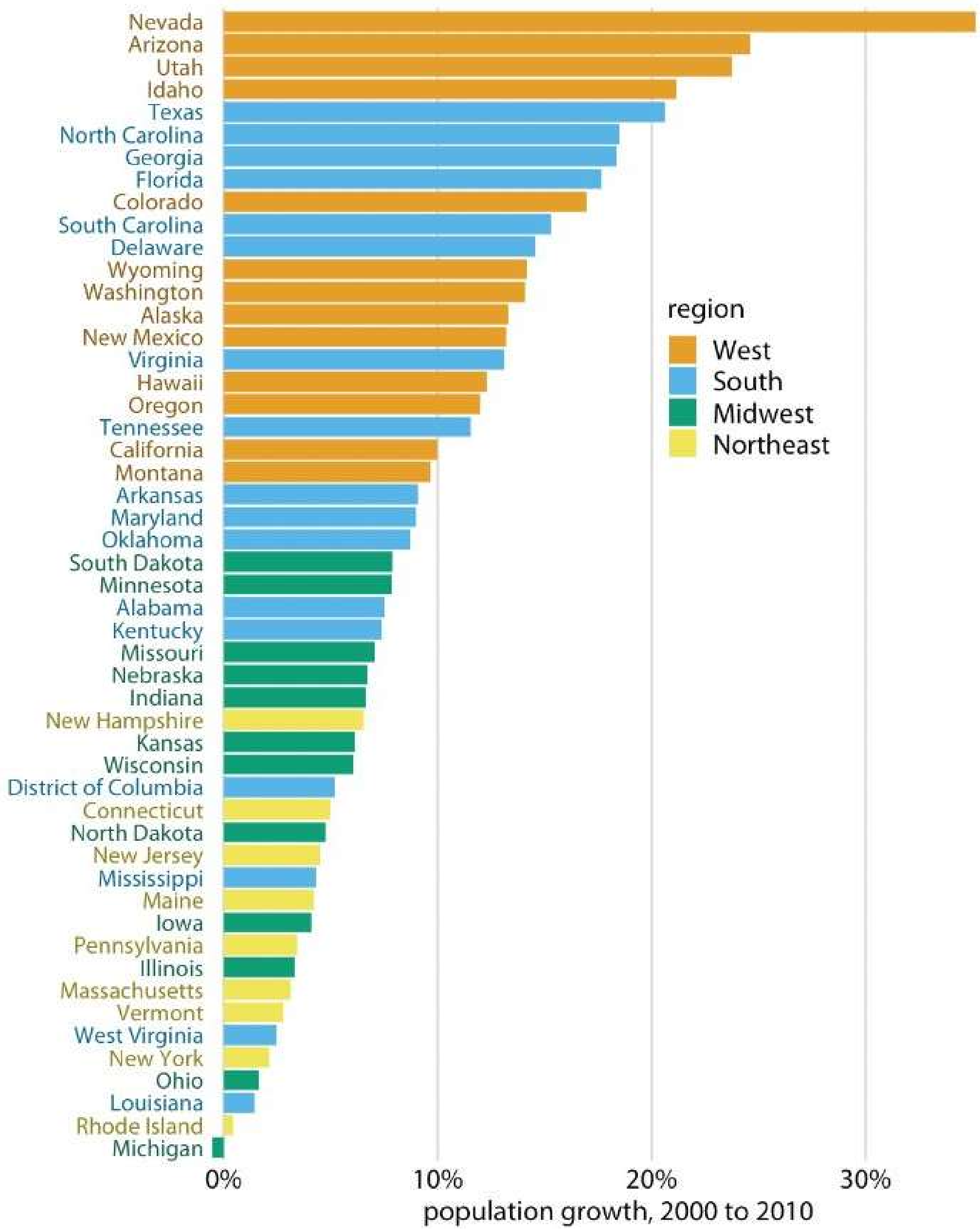
Fonte: Fundamentals of Data Visualization - Claus O. Wilke

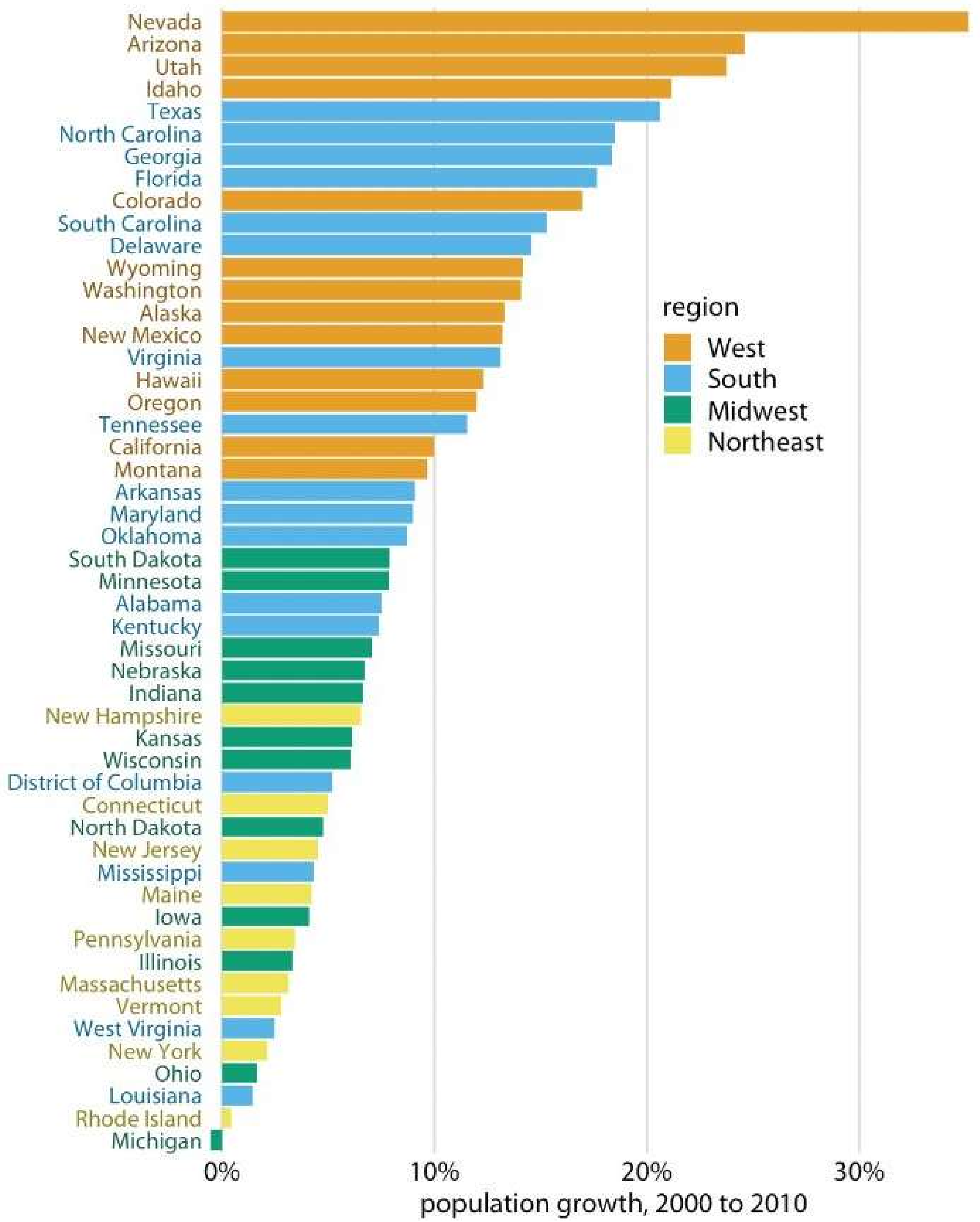
Coordenadas Cartesianas: X e Y

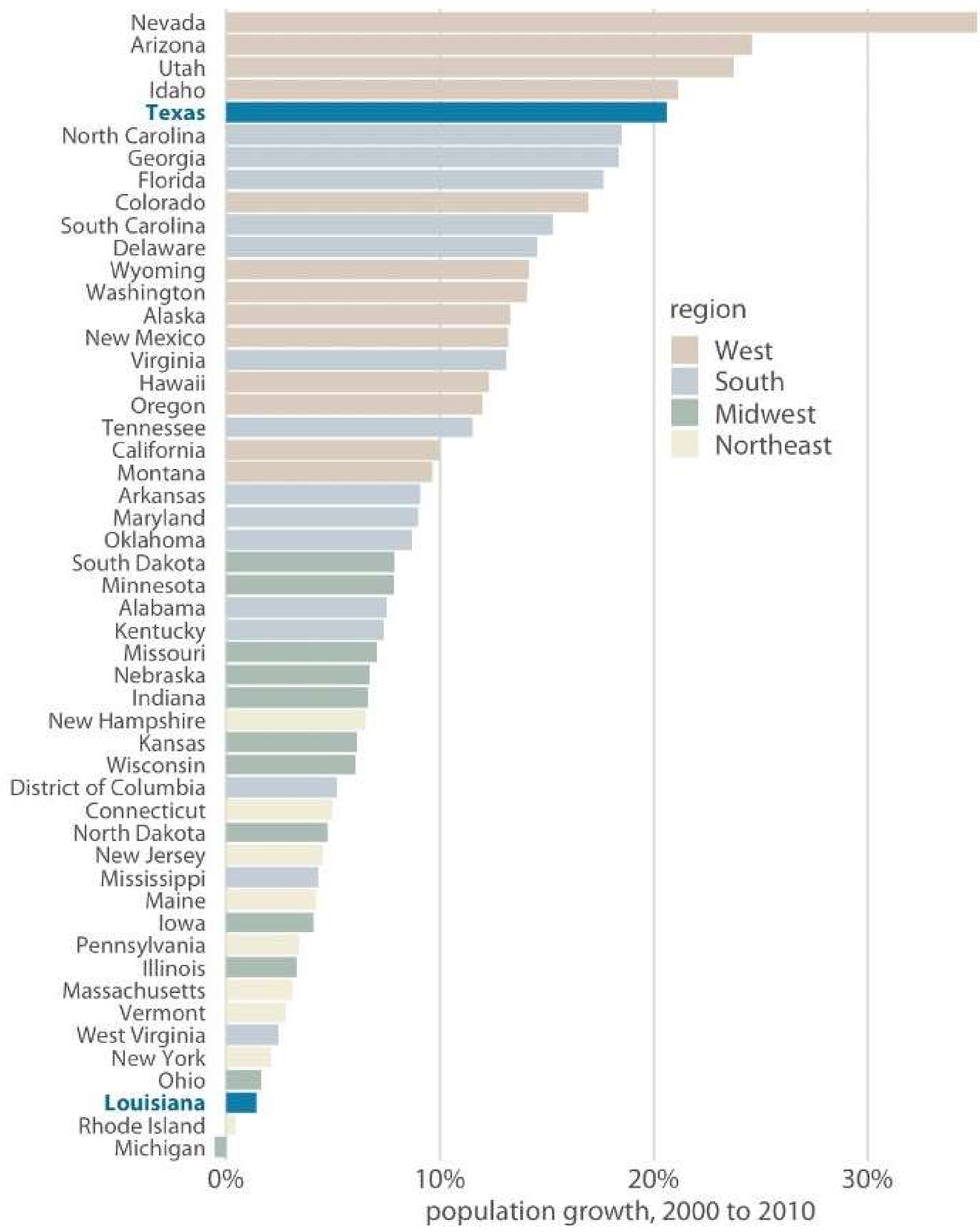


Cores




Nevada
Arizona
Utah
Idaho
Texas
North Carolina
Georgia
Florida
Colorado
South Carolina
Delaware
Wyoming
Washington
Alaska
New Mexico
Virginia
Hawaii
Oregon
Tennessee
California
Montana
Arkansas
Maryland
Oklahoma
South Dakota
Minnesota
Alabama
Kentucky
Missouri
Nebraska
Indiana
New Hampshire
Kansas
Wisconsin
District of Columbia
Connecticut
North Dakota
New Jersey
Mississippi
Maine
Iowa
Pennsylvania
Illinois
Massachusetts
Vermont
West Virginia
New York
Ohio
Louisiana
Rhode Island
Michigan
region
West
South
Midwest
Northeast
0%
10%
20%
30%
population growth, 2000 to 2010

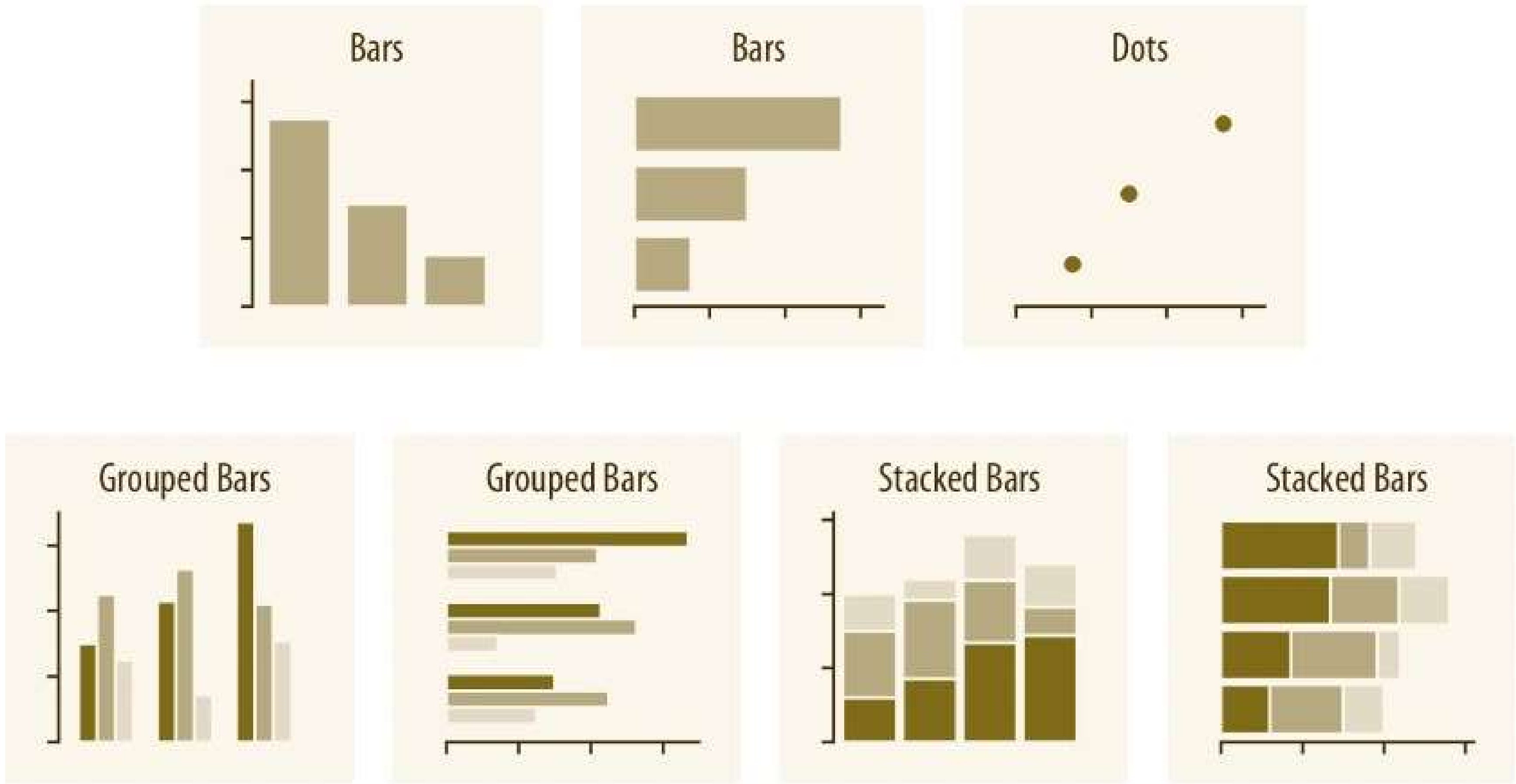

Nevada
Arizona
Utah
Idaho
Texas
North Carolina
Georgia
Florida
Colorado
South Carolina
Delaware
Wyoming
Washington
Alaska
New Mexico
Virginia
Hawaii
Oregon
Tennessee
California
Montana
Arkansas
Maryland
Oklahoma
South Dakota
Minnesota
Alabama
Kentucky
Missouri
Nebraska
Indiana
New Hampshire
Kansas
Wisconsin
District of Columbia
Connecticut
North Dakota
New Jersey
Mississippi
Maine
Iowa
Pennsylvania
Illinois
Massachusetts
Vermont
West Virginia
New York
Ohio
Louisiana
Rhode Island
Michigan
region
West
South
Midwest
Northeast
0%
10%
20%
30%
population growth, 2000 to 2010


Nevada
Arizona
Utah
Idaho
Texas
North Carolina
Georgia
Florida
Colorado
South Carolina
Delaware
Wyoming
Washington
Alaska
New Mexico
Virginia
Hawaii
Oregon
Tennessee
California
Montana
Arkansas
Maryland
Oklahoma
South Dakota
Minnesota
Alabama
Kentucky
Missouri
Nebraska
Indiana
New Hampshire
Kansas
Wisconsin
District of Columbia
Connecticut
North Dakota
New Jersey
Mississippi
Maine
Iowa
Pennsylvania
Illinois
Massachusetts
Vermont
West Virginia
New York
Ohio
Louisiana
Rhode Island
Michigan
region
West
South
Midwest
Northeast
0%
10%
20%
30%
population growth, 2000 to 2010

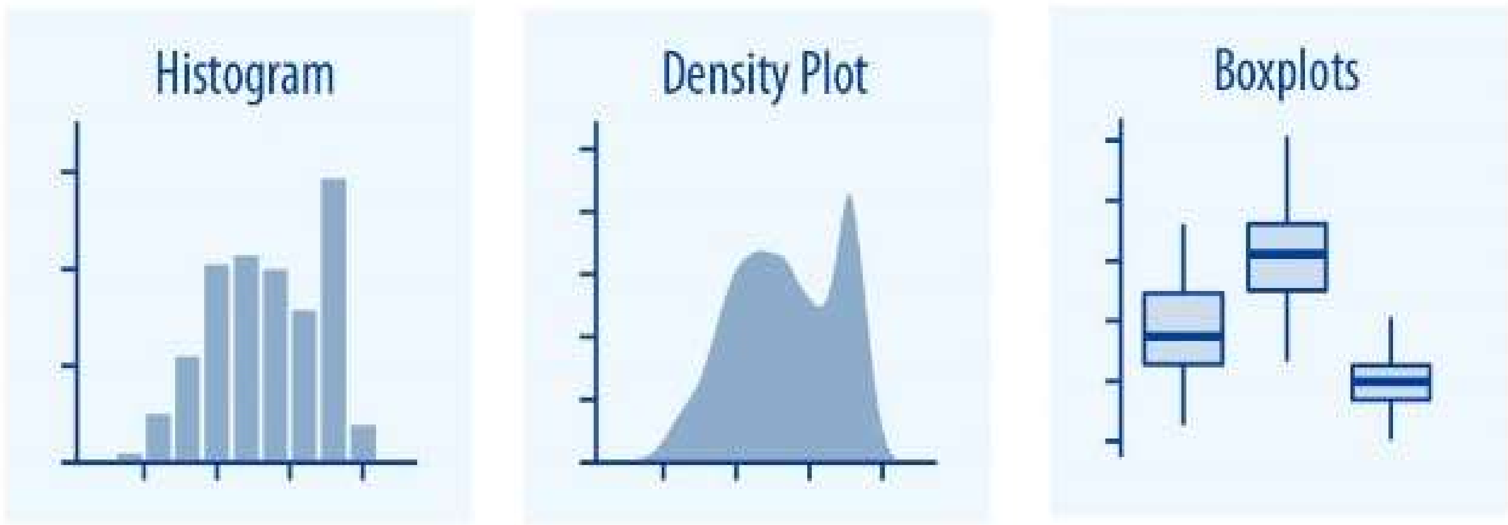
Diretrizes para visualização

Diferentes tipos de variáveis possuem visualizações tipicamente usadas para mapear seus valores em figuras gráficas

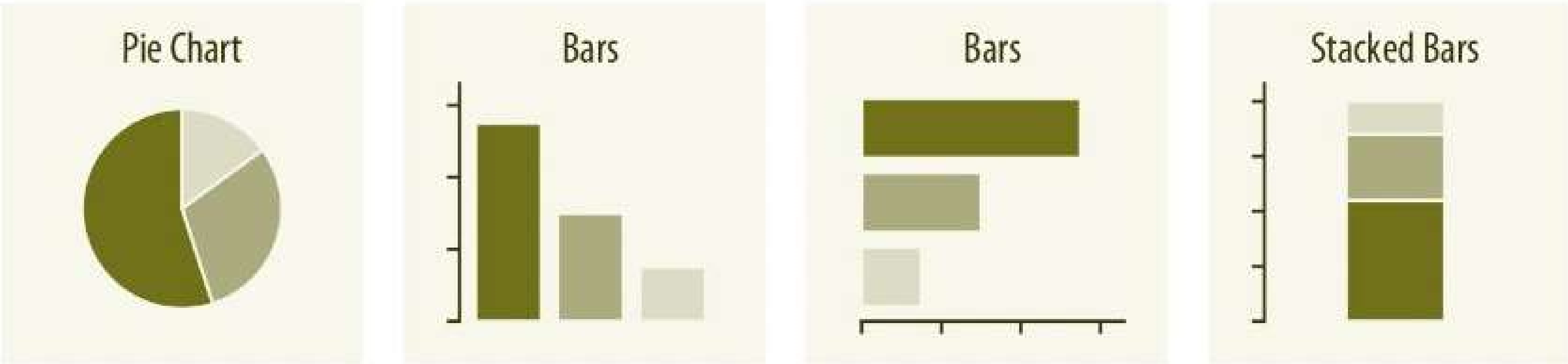
Quantidades



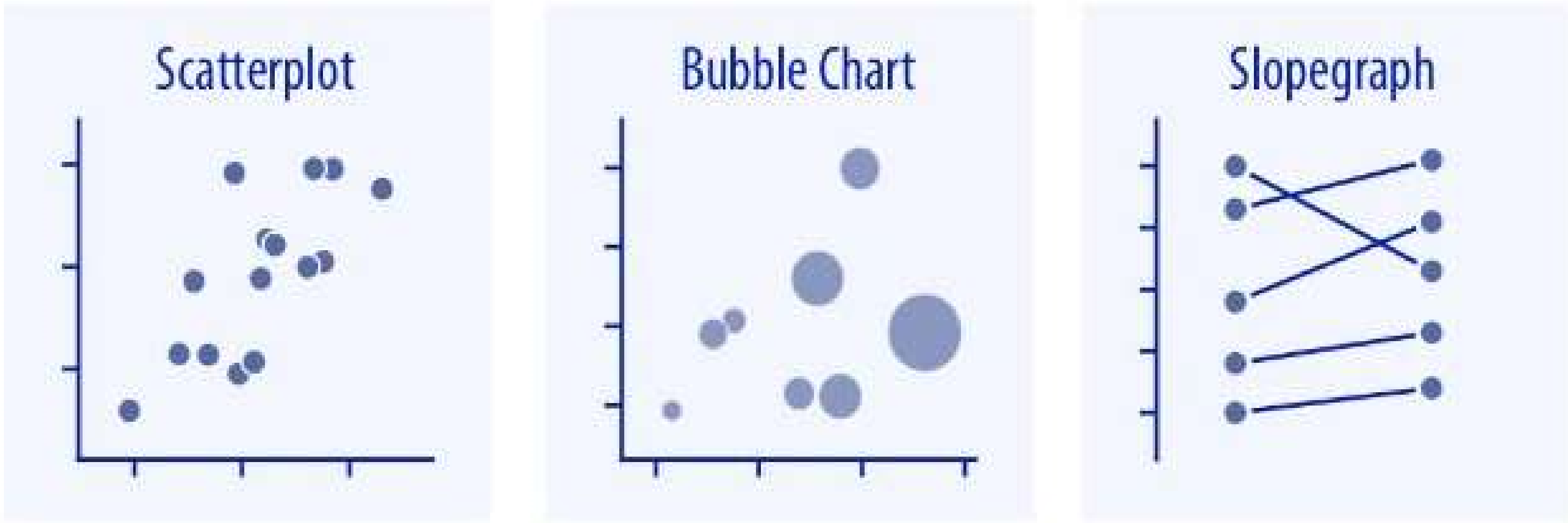
Distribuições



Proporções

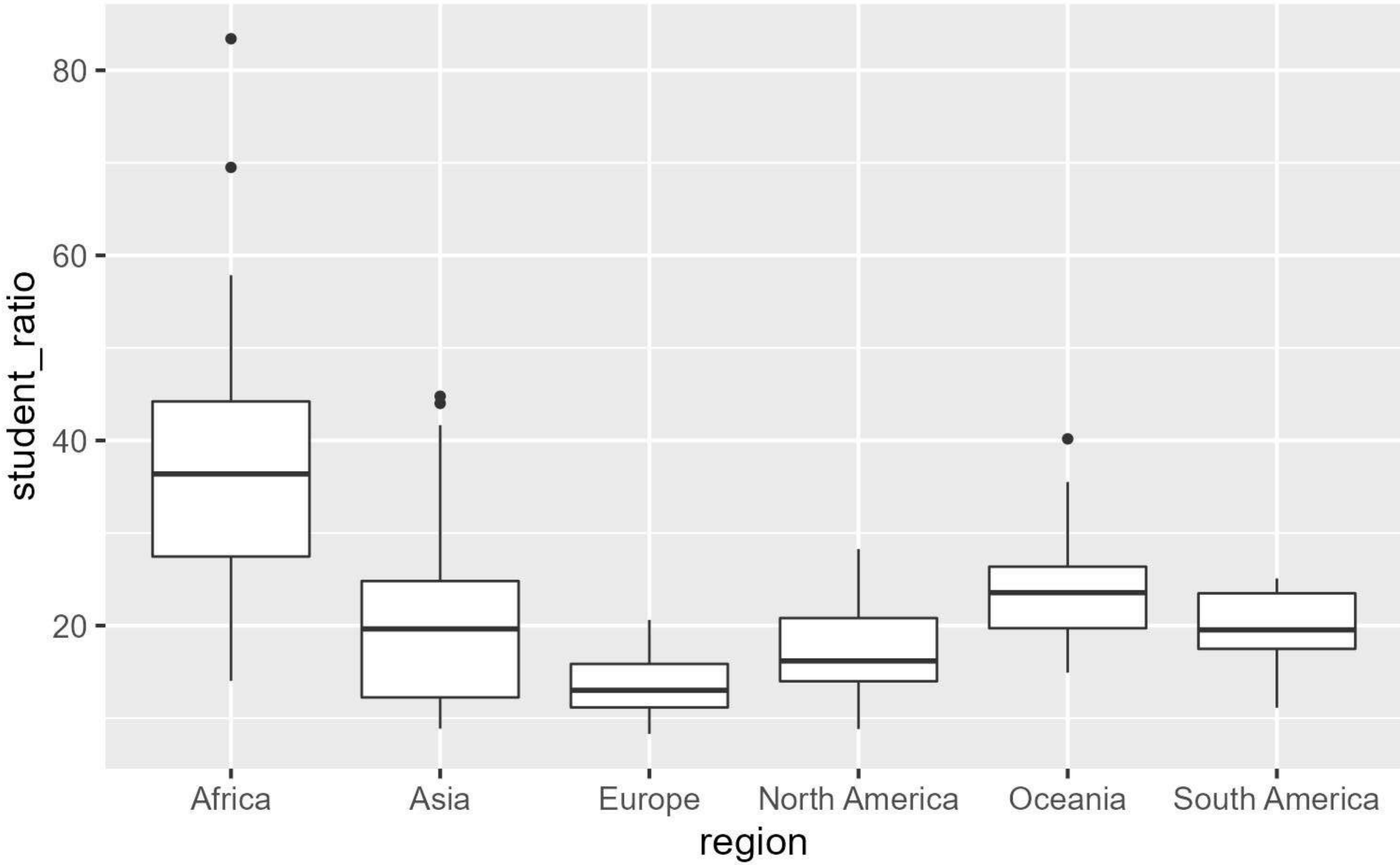


Relação X-Y

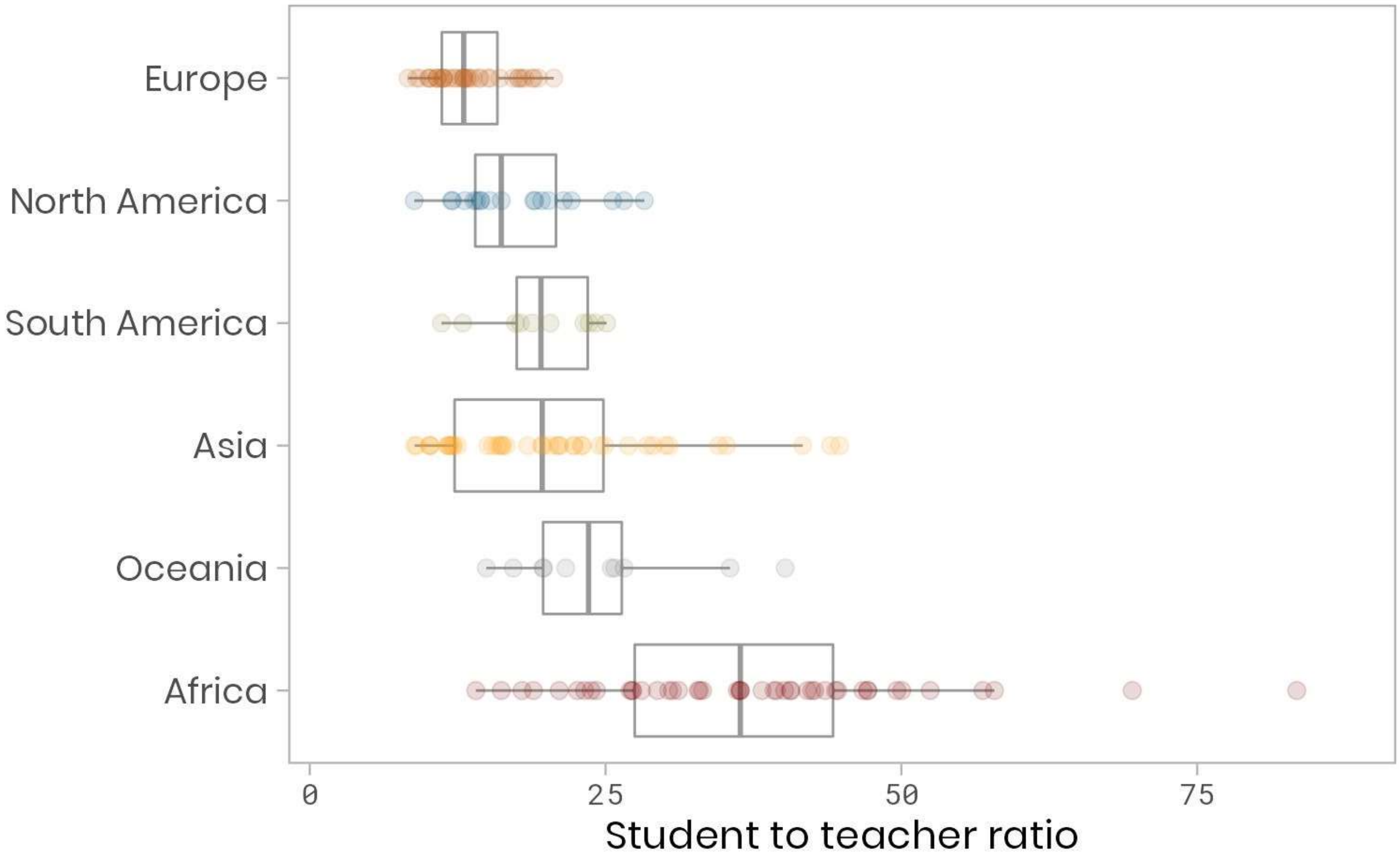


Vale a pena investir em visualização?

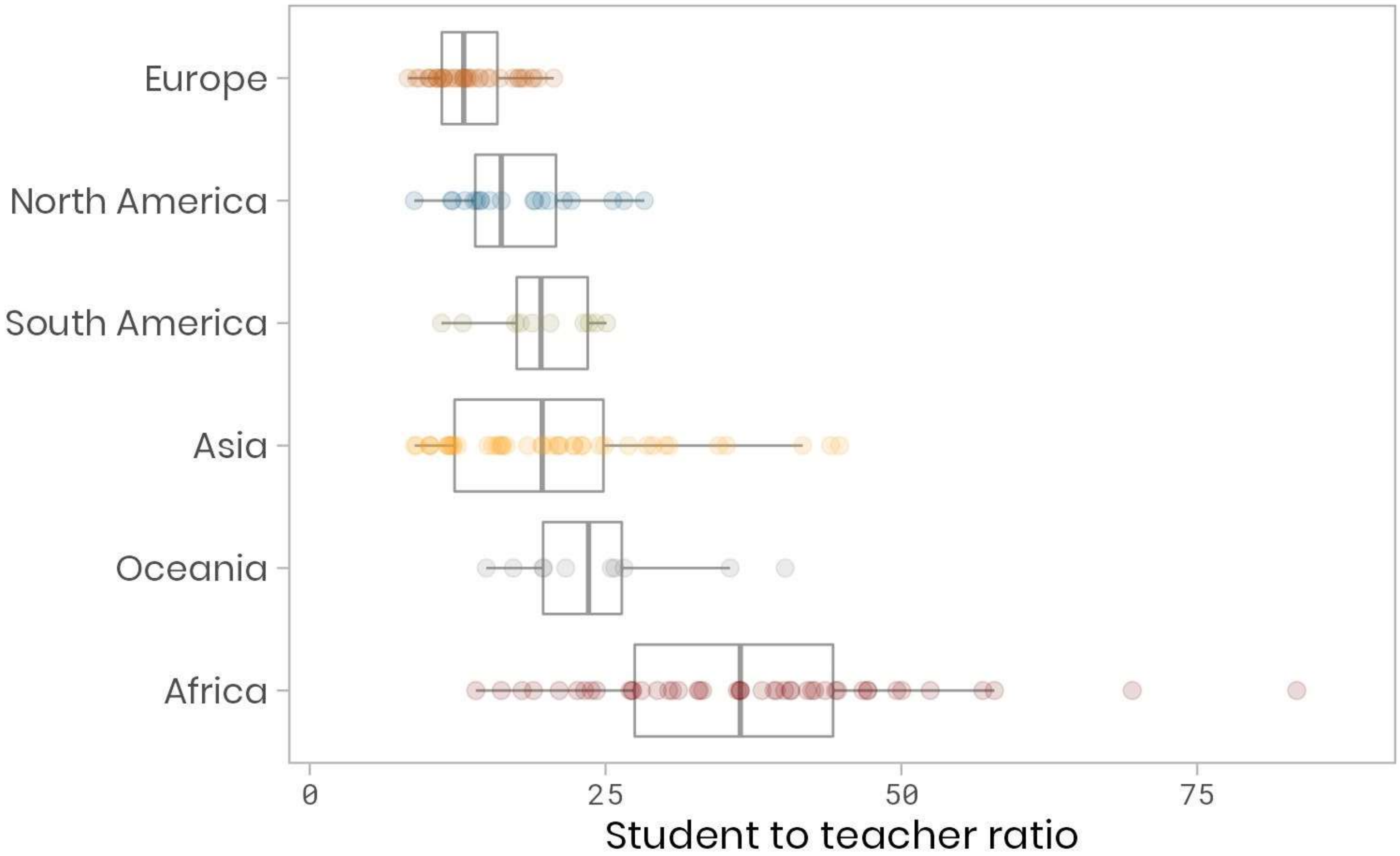
Relação Estudante x Profesor



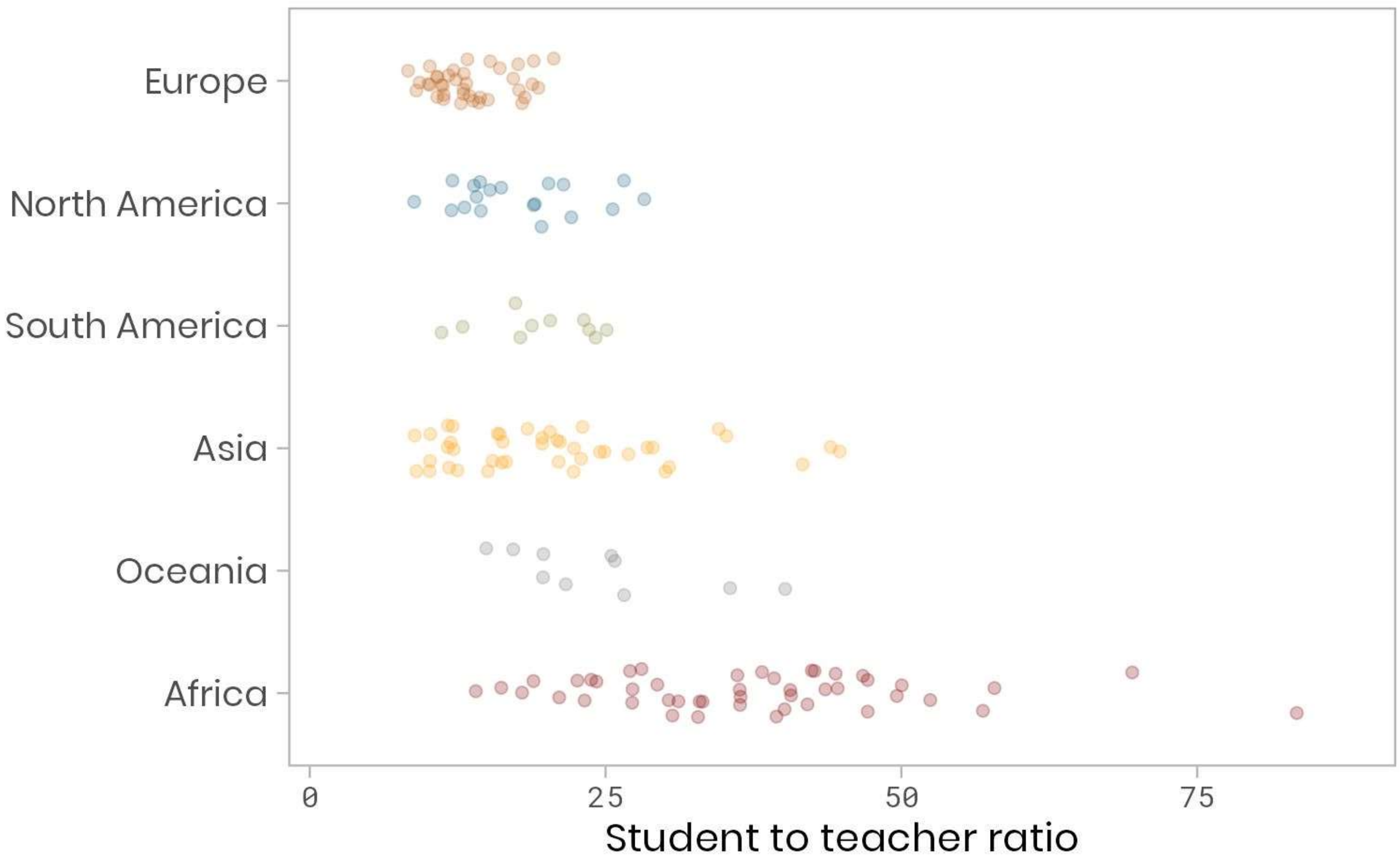
Relação Estudante x Profesor



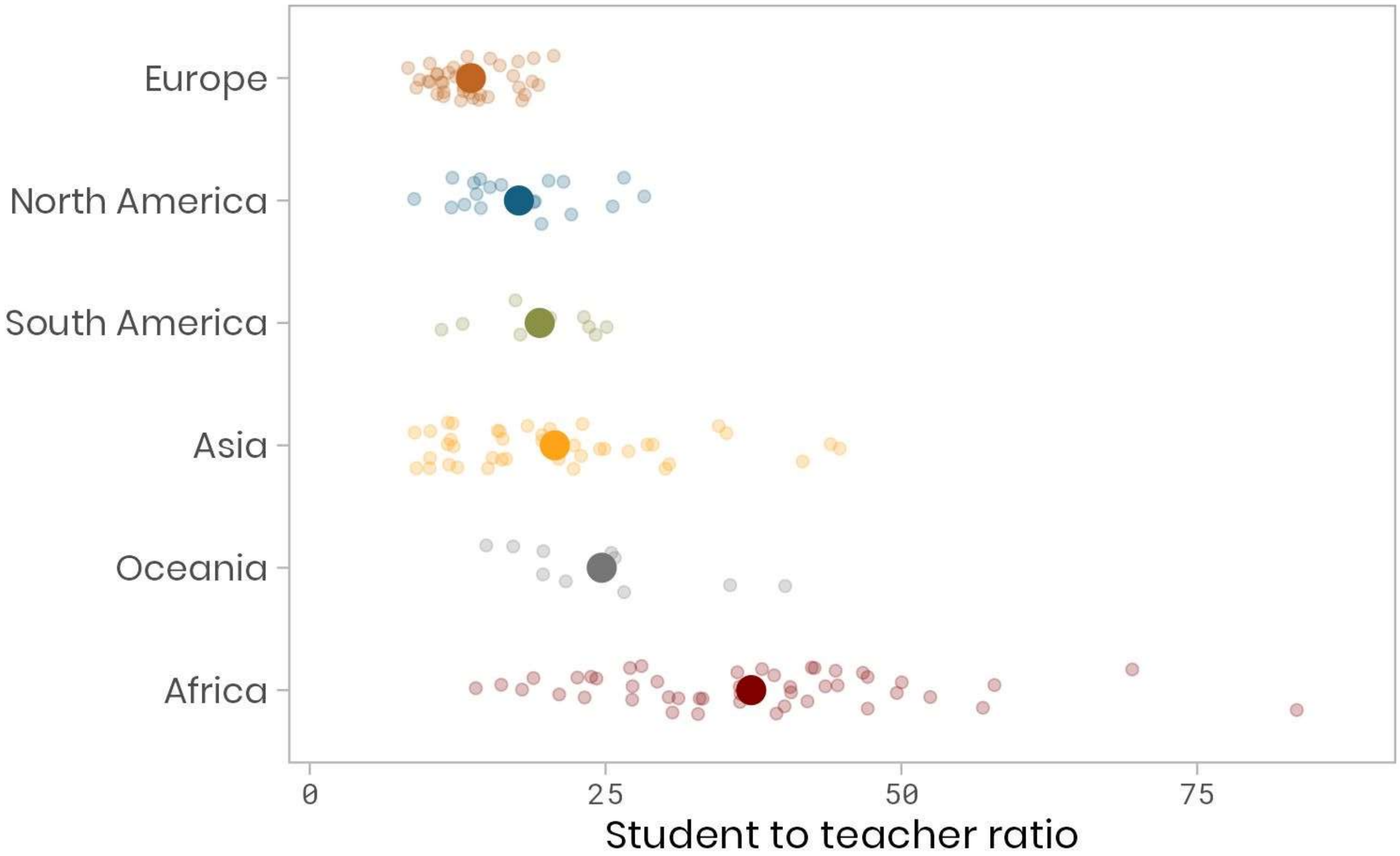
Relação Estudante x Profesor



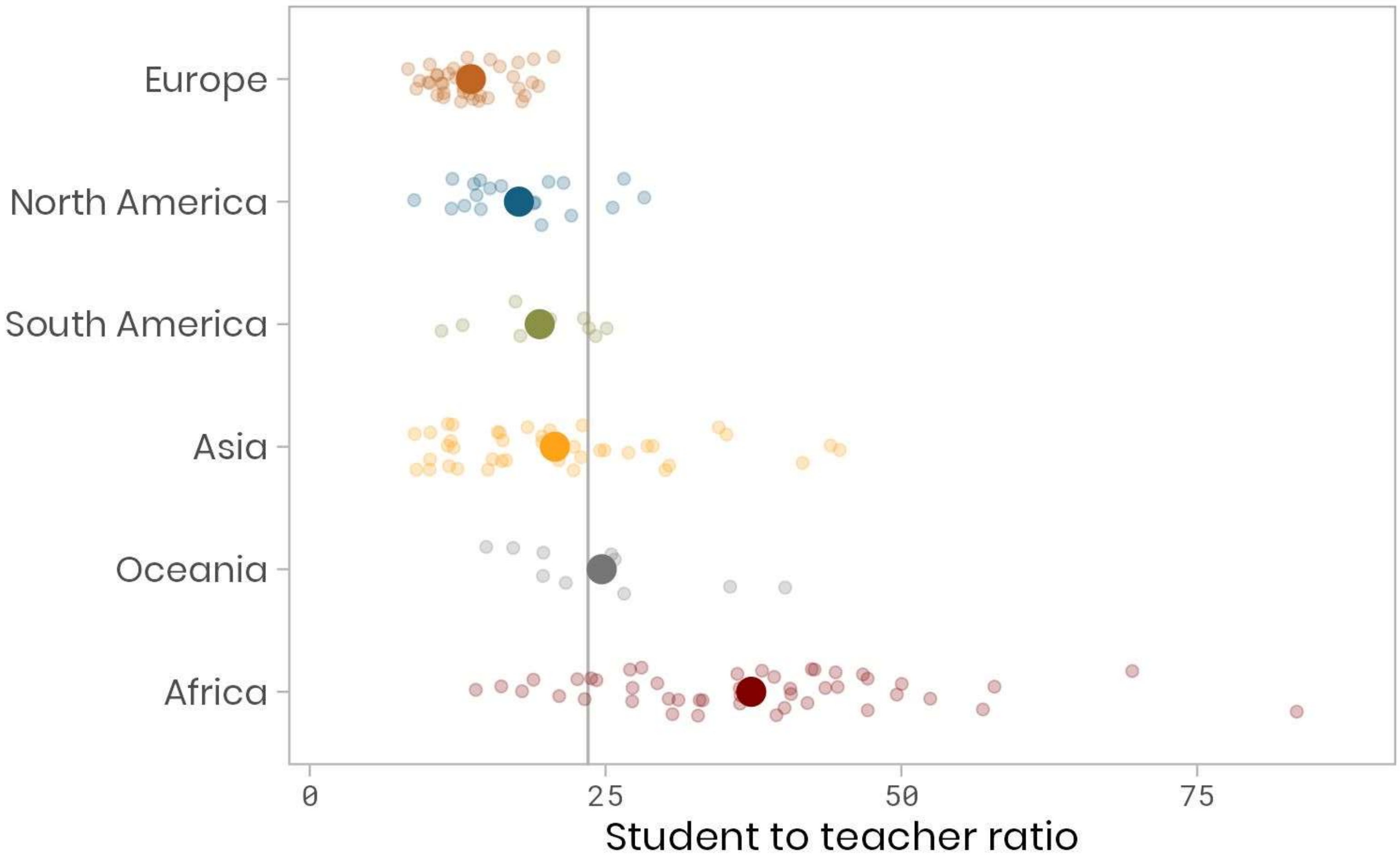
Relação Estudante x Profesor



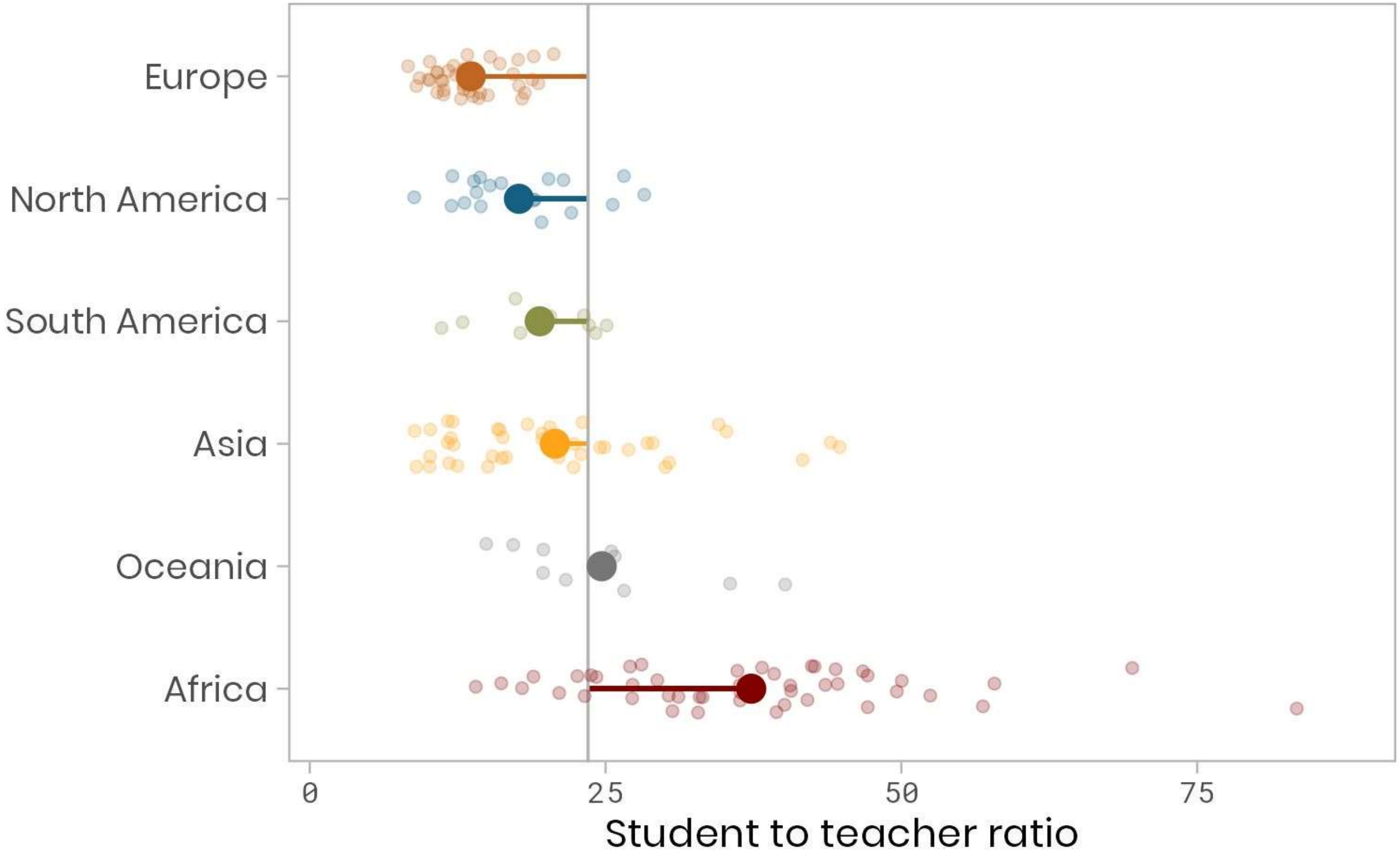
Relação Estudante x Profesor



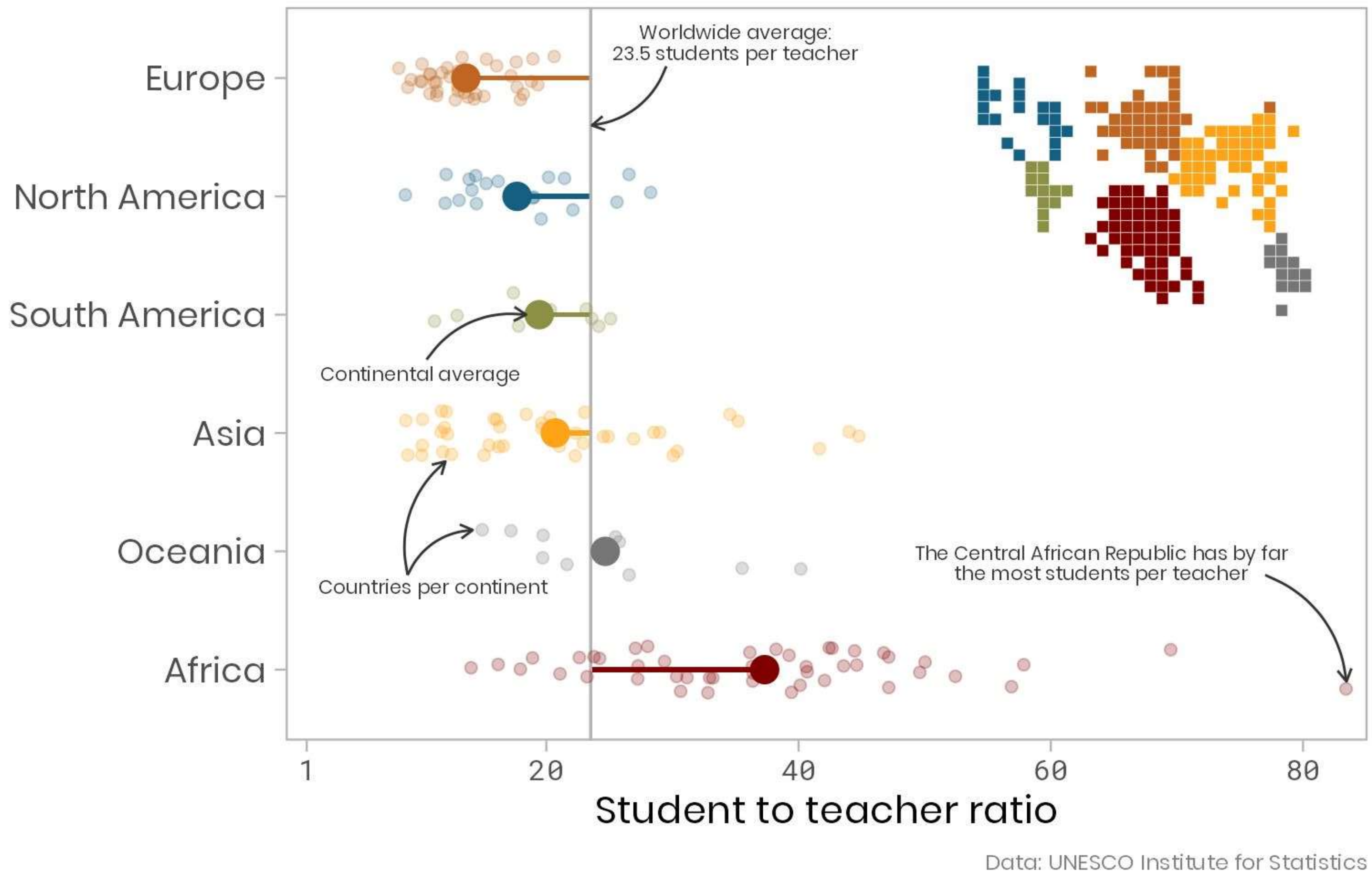
Relação Estudante x Profesor



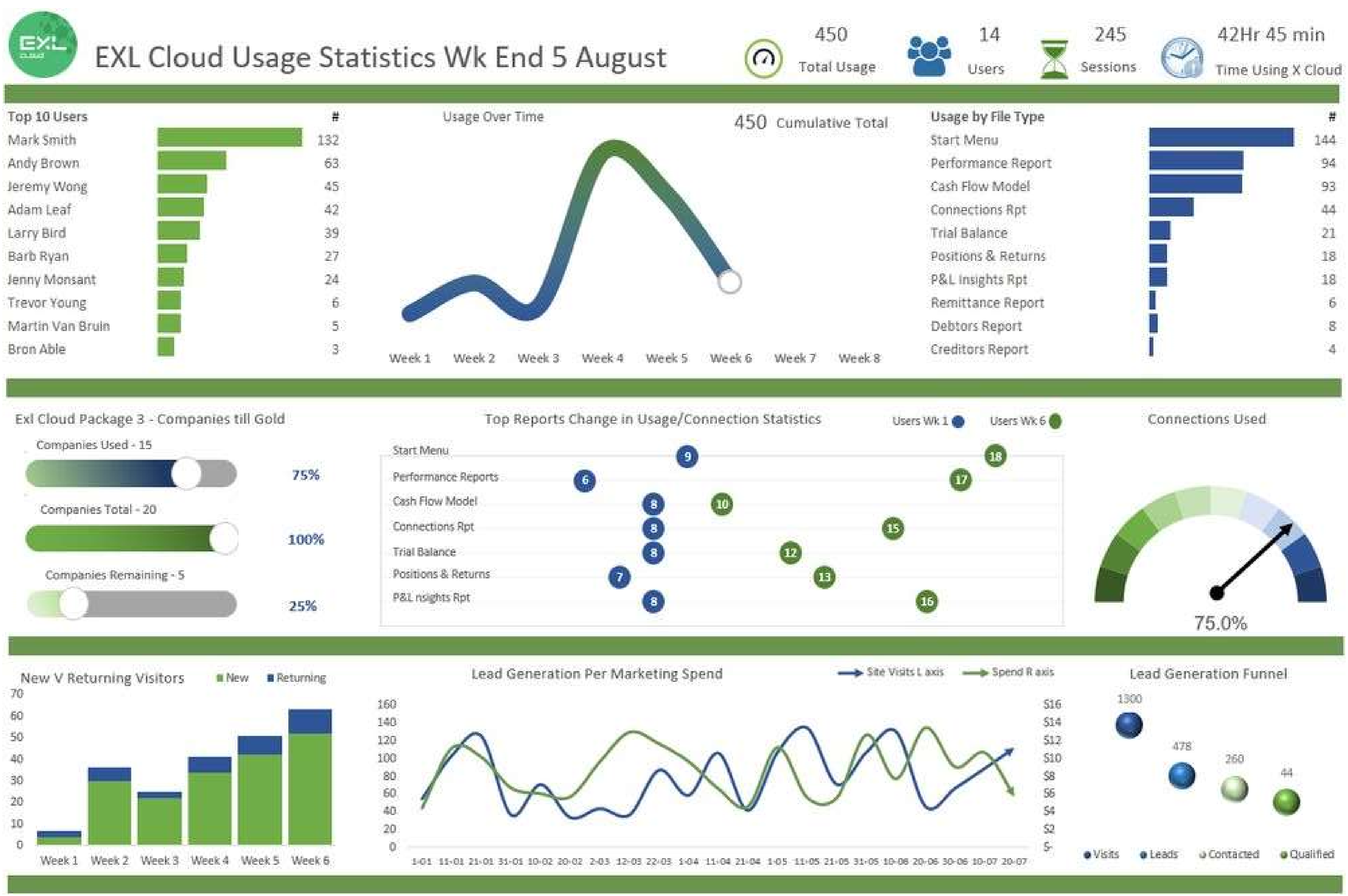
Relação Estudante x Profesor



Relação Estudante x Profesor



Dashboards


EXL Cloud Usage Statistics Wk End 5 August
450 Total Usage
14 Users
245 Sessions
42Hr 45 min Time Using X Cloud
Top 10 Users #
Mark Smith 132
Andy Brown 63
Jeremy Wong 45
Adam Leaf 42
Larry Bird 39
Barb Ryan 27
Jenny Monsant 24
Trevor Young 6
Martin Van Bruin 5
Bron Able 3
Usage Over Time
450 Cumulative Total
Week 1 Week 2 Week 3 Week 4 Week 5 Week 6 Week 7 Week 8
Usage by File Type #
Start Menu 144
Performance Report 94
Cash Flow Model 93
Connections Rpt 44
Trial Balance 21
Positions & Returns 18
P&L Insights Rpt 18
Remittance Report 6
Debtors Report 8
Creditors Report 4
Exl Cloud Package 3 - Companies till Gold
Companies Used - 15 75%
Companies Total - 20 100%
Companies Remaining - 5 25%
Top Reports Change in Usage/Connection Statistics
Users Wk 1 Users Wk 6
Start Menu 9 18
Performance Reports 6 17
Cash Flow Model 8 10
Connections Rpt 8 15
Trial Balance 8 12
Positions & Returns 7 13
P&L nsights Rpt 8 16
Connections Used
75.0%
New V Returning Visitors
New Returning
Week 1 Week 2 Week 3 Week 4 Week 5 Week 6
Lead Generation Per Marketing Spend
Site Visits L axis Spend R axis
Lead Generation Funnel
1300 478 260 44
Visits Leads Contacted Qualified